Acta Acustica united with Acustica
Vol. 104 (2018) 174 – 184

DOI 10.3813/AAA.919157

Non-Experts' Perceptual Dimensions of Voice Assessed by Using Direct Comparisons

Benjamin Weiss[1)], Dominique Estival[2)], Ulrike Stiefelhagen[1)]
[1)] Quality and Usability, TU Berlin, Ernst-Reuter-Platz 7, 10586 Berlin, Germany. BWeiss@telekom.de
[2)] MARCS, University of Western Sydney, Locked Bag 1797, Penrith NSW 2751, Australia.

Summary
Three data sets of 13 speakers each are analysed using the elicitation phase of the Repertory Grid Technique in order to identify vocal perceptual dimensions of non-expert listeners. Sentences read by female and male speakers of German (Phondat 1 corpus) and by male Australian English speakers (AusTalk corpus) have been rated on (dis)similarity by same-sex listeners using triples. Applying a balanced incomplete design proposed for the Repertory Grid Technique, frequencies of dissimilar pairs are transformed into distance measures using non-metrical multidimensional scaling. For both German data sets, three dimensions describing the speaker differences are found, whereas for the Australian data, we found four dimensions. By inspecting the individual labels given to extreme stimuli on each dimension, the authors subjectively named the respective dimension. Identical names for similar dimensions were avoided on purpose, aiming at reflecting each connotation best without inferring identity. The names are "calmness", "factual", and "naturalness" for German men; "tension", "positive timbre", and "maturity" for women; and "pitch", "remarkable timbre&voice", "emotion", and a fourth unnamed one for Australian men. Although an exact match cannot be supposed, there are strong similarities between dimensions from the different data sets (factual/tension/emotion, naturalness/maturity, timbre/timbre&voice). A smaller pretest supports these findings. Overall, the dimensions observed are found to be more complex and person-related than typically described in literature. These results add to the current state of research in perceptual dimensions of non-experts, and represent a foundation for developing a questionnaire to assess listeners' impressions.

PACS no. 43.71.Bp

1. Introduction

The impression of speakers' voices has fascinated people for a long time. Although people try to estimate relevant speaker traits and states when listening to unacquainted persons, describing the voice itself, which carries such information, is a topic of its own. As the first impression is typically assumed to be important for estimating the social relevance of the speaker, the vocal attributes causing this impression of person might be unconscious until explicitly asked for. But still, descriptions of voice quality and speaking style are frequently referred to – and thus made conscious – for instance in newspaper reports on poli ticians' speeches or even in literature. In paralinguistic research, e.g. on externalization and attribution of personality or affect, or on inter-personal relationships, such vocal descriptions are used to represent listeners' perception. By following the Brunswikian lens model [1], such ratings are applied to systematically study the relation between listeners' attributions and acoustic features [2, 3].

Received 24 October 2016,
accepted 15 November 2017.

However, there is no consensus on the dimensions commonly perceived when confronted with unacquainted speakers. For trained voice professionals, schemata do exist to categorize and label for example voice quality and articulatory settings [4] or speech pathologies [5]. For non-experts, who lack profound knowledge on articulation, however, various different questionnaires or rating schemes have been proposed to collect user impressions in experiments [6, 7, 8, 9, 10], typically stimulated by natural inter-speaker variation like a low versus a high pitched voice. Despite such data and instruments available, the common ground is not very satisfying. This is especially true for assumed dimensions of voice quality and timbre (cf. Section 2).

According to Laver [11], impressionistic labels typically used by non-experts are holistic, often refer to similar impressions, and thus exist in large numbers, compare e.g., "warm–cold", "full–thin", and "dark–bright". Critical issues arising from the usage of such labels in pre-defined questionnaires concern the possibility of not representing perceptual dimensions well or that the interpretation of items depend on grouping and ordering [12].

Accordingly, the difference between voice experts and non-experts originates in perceptual training, knowledge

of articulatory processes and an agreed upon naming scheme for labelling. An approach circumventing the issues of pre-defined questionnaires for non-experts is to collect listeners' impression of speakers without labels at all. This is usually done by directly comparing stimuli to collect similarity measures, which are analysed subsequently by multidimensional scaling (MDS) to extract the most relevant dimensions. Prominent examples are studying sound impressions [13], speech pathologies [14], acoustic correlates of voice quality differences [15], and speech transmission degradations [16]. As a limitation of this approach, the resulting dimensions are difficult to interpret. Also, collecting such measures is a very time-consuming method compared to questionnaires, as a large number of combinations of speakers' stimuli have to be presented, and this method is usually limited to only the most salient feature(s).

Vocal dimensions might be highly culturally dependent, as language or peer-group specific characteristics may be perceived as quite salient for one group of listeners, but barely noteworthy for others. As an example, the recent trend of young, "upwardly mobile" females in the U.S. to increasingly produce creaky voice [17] is considered as very familiar to younger people with a similar social background. However, it is highly salient (and evaluated negatively [18]) for other groups in the US, for example, older adults [19, 20].

After presenting related work in the following section, a pre-test and three studies are reported (two with male and one with female speakers).

2. Related Work

There is no coherent body of results concerning perceptual dimensions for spoken voice characterizations. Although the ranking of such dimensions will naturally depend on speaker selection [21], the dimensions found so far apparently do also depend on the kind of stimulus (segment, sentence, and paragraph), as e.g. single vowels emphasize the relevance of formant frequencies and lack temporal information [22]. Also, speech pathologies [23, 14] and gender [23, 22] constitute major perceptual categories, and perceptual dimensions may be different for male and female speakers or listeners. In this section, we concentrate on sentences or paragraphs as stimulus material, and neglect smaller segments for the sake of validity of global dimensions.

Most studies aiming at perceptual dimensions of speakers' voices have used questionnaires, subsequently analysing the data with factor analysis. As the dimensions found are not consistent, results of relevant studies are presented here successively, starting with studies applying questionnaires before reporting those studies that apply similarity measures.

2.1. Studies applying questionnaires

In an early study with 32 subjects, four dimensions have been reported for 16 male speakers uttering sentences [24]. These are 1: *clarity*, 2: *roughness*, 3: *magnitude*, and 4: *animation*. The gender of the listeners has not been reported.

Using visual analogue scales to assess vocal impressions only, paragraphs uttered by 35 male teachers and 36 male actors in a normal loudness condition and judged by (prospective) voice therapists resulted in four dimensions, named as 1: *variation and sonority*, 2: *irregularity*, 3: *degree of noise*, and 4: *phonatory effort* [25].

In [26], the dimensions 1: *relaxed*, 2: *dynamic*, 3: *dark*, 4: *unremarkable*, 5: *professional*, and 6: *fluent* are reported, based on data from 46 participants rating five male and five female speakers' utterances.

An earlier approach similarly employed stimuli of five male and five female speakers and a comprehensive questionnaire, but presented them to 235, mostly female listeners who were students of, e.g., speech therapy. This study yielded 5 dimensions: 1: *melodiousness*, 2: *articulation quality*, 3: *voice quality / timbre*, 4: *pitch*, and 5: *tempo*. Dimension 1 related to voice appreciation, and Dimension 2 related to perceived self-confidence [27].

Scherer [9] assessed vocal percepts of one-minute collections of three 20 sec. samples from a jury discussion. The 31 German and 28 US-English male speakers recorded, each had to rate their own samples and let them be rated by three peers of their own choice. The voice qualities *pleasantness, resonance, depths, breathiness, warmth, thinness, pitch*, and the speech characteristics *precision, rate, sloppiness, rhythm, affectation*, and *regularity* were found.

2.2. Studies applying direct comparisons

In their comprehensive first chapter, Kreiman and Sidtis [21] conclude that there are no satisfying results for "independent and valid" factors for voice characterization and they stress the important finding of listeners'/raters' idiosyncrasy as reported by Voiers [24]. They argue that there seems to be no "common perceptual space" for listeners based on results presented of 80 male and 80 female speakers being rated separately on dissimilarity in pairs by eight experts each on the basis of vowel recordings [14]. However, their stimuli included highly pathological voices and used sustained vowels as material. As a result of individual dimension reduction, speakers were rather placed in highly individual clusters in the dissimilarity space, than spread out evenly and consistently. Subsequently, a conceptual framework was compiled in order to increase reliability for expert voice assessments for clinical purposes, taking into account for example the inclusion of external references [28]. For our purpose of studying listeners percepts instead of clinical diagnostics, analysing reliability is likewise important, but the reported degree of, e.g. intra-class-correlation (ICC) between .55–.85 is considered as sufficient concordance for further calculation, i.e. aggregating subjective data for dimension reduction.

The study by [14] is a typical example of the identification of perceptual dimensions based on dissimilarity measures for pairs of stimuli instead of rating each stimulus separately on a questionnaire. Their use of sustained

Table I. Dimensions found in related work and mapped to categories. [1]: *Masculinity* may also be assigned to effort or timbre. [2]: With the meaning of swift. [3]: This dimension reflects "brightness" and is therefore better assigned to Timbre instead of Voice Quality.

Categories	Terms used for dimensions/factors
Average Pitch	*pitch* [27, 22, 15, 9, 31], *masculinity* [30][1]
Intonation	*melodiousness* [27], *variability* [30, 26], *sloppy* [9]
Rhythm	*rhythm* [9], *regularity* [9], *fluent* [26]
Tempo	*tempo* [27], *duration* [22, 31], *rate* [9], *animation* [24], *dynamic* [26][2]
Pronunciation	*articulatory quality* [27], *precision* [9], *clarity* [24], *professional* [26]
Timbre	*voice quality* [27][3], *magnitude* [24], *dark* [26], *depth, warmth, thinness, sharpness, gloom, flatness, dryness, nasality* [9]
Voice Quality	*creakiness* [30], *hoarseness* [22, 9], *unremarkable* [26], *resonance, roughness, breathiness* [9]
Vocal Effort	*effort* [22, 31], *harshness* [9], *loudness* [9], *relaxed* [26]
others	*affectation* [9]

vowels as stimuli aligns with the majority of studies using paired comparisons (cf. [29]). And typically, too, acoustic correlates of speaker distances are provided as potential acoustic causes for listeners' impressions. Such acoustic correlates may be considered only as a first hint for identification.

However, there are also some studies using sentences or phrases instead of vowels. For sentences for example, five dimensions were found in [15]. They asked ten male and ten female listeners to rate 15 male voices on a 9pt scale. The dimensions correlated with acoustic measures of 1: average fundamental frequency (F0), and 2–4: the formants F3, F2, and F1 of the vowels, respectively. For Dimension 5 no correlation could be found.

For whole passages read by 20 males, the dimensions 1: *pitch and effort* and 2: *hoarseness* were found for ten, mostly female, experts [22]. The third dimension correlated with the difference between F2 and F1, whereas the fourth dimension could not be interpreted. A similar solution for 20 females is reported as 1: *effort and nasality*, 2: *pitch*, 3: *duration*, 4: no interpretation.

In the last study to be reported here, 24 participants rated (dis)similarity of stimuli from 22 male speakers [30]. Based on expert opinion of five phoneticians, and additional acoustic measures, the resulting four dimensions were named 1: *perceived masculinity*, 2: *creakiness*, 3: *variability*, and 4: *mood*. The first dimension is, amongst others, related to "pleasant", "low", "relaxed", and "heavy" attributes.

2.3. Summary

From the selection of studies presented, it becomes evident that the experimental conditions of speaker's gender, speaker's regional background, and speech pathologies are usually controlled for, as these aspects are known to be attributed to voices and therefore are quite consistently included in the rating schemes [31, 32]. Before summarizing the remaining dimensions in detail, other aspects of the presented studies have to be commented on: Although fluency is perceived by listeners and consequently used to attribute proficiency [33], speakers' expertise in speaking relating to the impression of fluency and pronunciation precision is not always reported. Exceptions from this are Bele [25], who recruited teachers and actors to ensure variability in proficiency by design, and Voiers [24], who implied to have recruited non-professional speakers. The gender of listeners does not seem to play an important role [27] and also is not always reported. Another issue with some of the studies is that participants often are students of relevant speech related disciplines, and thus deviate from mainstream naïve listeners in at least having an inclination for the field. Also, there is an imbalance towards male voices. Lastly, there are implicit and explicit differences in the definitions of the actual aspects to be studied: Some studies apparently aim at (phonetically defined) voice quality, others at the "common" voice, i.e. voice, speaking style, and pronunciation [27]. Like the latter, the authors want to study the broad perceptual impressions of non-experts, and choose sentences for material, while controlling for obvious disfluencies.

Whereas for pathological voices it was hard to find common perceptual dimensions [21], there are dimensions for "normal" voices (here defined as not deliberately pathological ones) frequently occurring based on this rich, but quite diverse body of results. Also, consensus between listeners can be regarded as sufficient for at least some perceptual dimensions. Thus, the aim is to distinguish such percepts with reasonable consensus from idiosyncrasies. To do this, Table I summarizes the individual studies by categorizing the dimensions found and by listing the number of occurrences of dimensions. The reader might disagree on individual decisions of mapping assumedly related results to the same category, as these are not empirically founded. For example, instead of considering harshness as result of Vocal Effort, it also might be subsumed with other aspects of voice quality. Likewise, resonance might also fit to Timbre [34]. But these categories represent an abstraction process during the work on related studies and support the comparative analysis of their results.

In the following overview, two studies that have been mentioned above are excluded because they applied expert schemes on the basis of well-defined phonetic dimensions, instead of recruiting non-experts [14, 25]. As most studies relied on male speakers the dimensions found may not be generalized with respect to gender.

Based on the data illustrated in Table I, we argue that the prosodic features of Average Pitch and Tempo represent established perceptual concepts. Other frequently occurring dimensions are related to the categories Pronunciation (articulatory precision), Vocal Effort, Intonation (melodiousness/activity), Voice Quality (laryngeal settings), and Timbre.

It becomes obvious that there is a mismatch in complexity: The first dimension found in [30] is named *masculinity*, for example, and is actually an evaluation related to a "relaxed", "dark" and "low-pitched" voice of male speakers. A connectivity between a "low" voice (Pitch) and a "dark" voice (Timbre) has also been found [26] and can be physiologically grounded in cases of other variables kept constant (e.g. muscular tension). Another example for such an evaluative dimension is Scherer's first factor pleasantness including e.g. "clear" and "melodic", although there are separate factors of precision and sloppiness [9]. In contrast, Timbre is a pparently a superordinate category for multiple perceptual impressions.

As a last comment it shall be noted here that the category of Voice Quality comprises conflicting results: The questionnaire labels "breathy", "creaky", "hoarse" (and "nasal") load on a single factor in [26], leading to a general impression of remarkable voices, while the factors of *hoarseness, roughness* and *breathiness* could be distinguished in [9].

In order to resolve the conflicting and unclear results, especially for the categories Timbre, Voice Quality, and Rhythm, this paper presents new empirical evidence for perceptual dimensions. Following the introductory critique on long item sets in questionnaires, a method of direct comparison has been chosen. Enriching this approach by asking for individual, subjective labels a posteriori provides first indicators to identify the extracted dimensions and to choose appropriate acoustic measures at a later stage. To identify which perceptual dimensions occur frequently, we applied this method on four data sets:

As the first set has already been presented elsewhere [35], the results are described only briefly in this section. Also, the subsequent improvements in the procedure and design for the main experiments are explained here. It is a rather small data set of German speech that illustrates the procedure and the dimensions identified support the findings of the main experiments. The three larger studies are presented in detail later on: two German ones (males and females), and also one with male Australian English speakers for a cross-linguistic comparison.

From a public database, nine sentences from five male and five female speakers were rated on binary similarity in triples with identical sentences within each triple, and applying a complete design (i.e. all combinations of the five speakers for each gender). The sentences were randomly assigned to each triple to avoid boredom. For these ten speakers, the two resulting dimensions for each gender correlate strongly with factors for the same speakers, but different listeners, obtained from a questionnaire, namely 1: *likeability and activity*, 2: *attractiveness* for male speakers, and for female speakers Dimension 1: *likeability and dominance*, and Dimension 2: *proficiency*. Labels were not analysed for naming of the dimensions. The most striking result is that only *proficiency* relates to voice and speaking style descriptions and comprises for example pronunciation precision, whereas *likeability, attractiveness*, and *activity* relate to speaker attributes that have been assessed on a separate part on the questionnaire. Either the listeners separated the speakers on their person attributions inferred by acoustic/phonetic features, or those acoustic/phonetic features that are relevant and correlated to the person attribution were not covered properly by the questionnaire. The dimensions obtained are certainly not exhaustive, as there were only five speakers presented for each sex. One undesired, but frequently occurring class of free text labels was related to sentence stress positioning. This was considered as database artefact due to the isolated reading condition during recording of the database. As a consequence, the procedure was modified for the following studies by a) recruiting same-sex listeners to potentially reduce the occurrences of labels related to attractiveness, b) advising the participants to neglect sentence stress positioning, and c) analysing the labels for dimension naming instead of conducting a separate experiment with an external questionnaire.

3. Procedure of the main studies

The method applied is the Repertory Grid Technique used to analyse personal constructs [36], now widely used, e.g. for characterization of sounds [37], stuttering [38], or user interface evaluation [39]. The first phase of such a Repertory Grid Technique is called the elicitation phase, as it uses the binary similarity ratings of triples of stimuli to elicit individual descriptors representing the constructs. This elicitation phase has also been applied for all four studies presented here: Triples of stimuli from separate speakers are presented to the listener, who is asked to identify the pair of voices (stimuli) most similar to each other, in contrast to the third (dissimilar) stimulus. After that, the listener must describe in her own words (free text field) the similarity for the chosen pair (e.g., both "fast"), and how that pair differs from the third stimulus (e.g., "slow"), thus eliciting a pair of (assume d) opposite attributes for each triple (Figure 1). Furthermore, the listeners had to indicate whether the first attribute of the pair was considered negative, positive, or neutral.

The assumed benefit of such triples, compared to a paired stimulus presentation with a distance measure, is its independence from external references, i.e. expected variety in speakers or order of presented pairs, since the reference is provided within the triple. Based on the experience from the pre-test, listeners were told to ignore sentence stress differences, as the stimuli were read. They also were asked to avoid labels referring to speaker traits or states, such as emotion, attitude, or age. Instead, they were asked to concentrate on voice and speaking style.

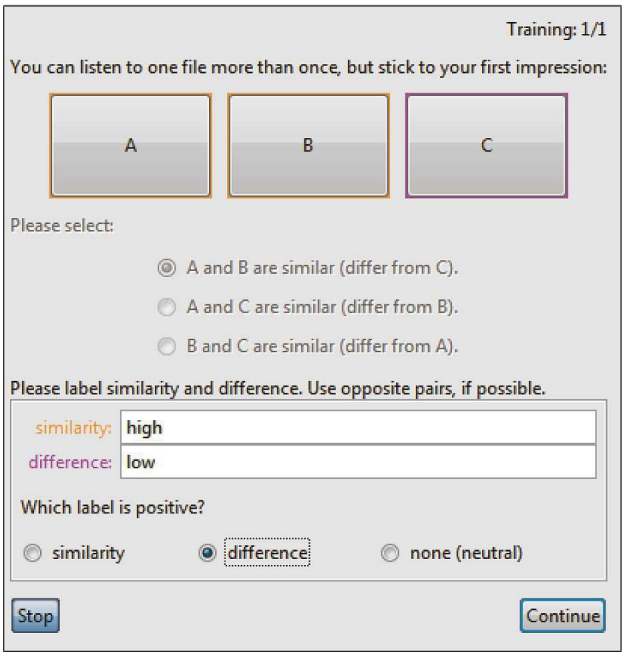

Figure 1. Screenshot of the experimental interface.

As the experiment was designed to last one hour in order to avoid fatigue, a balanced incomplete block design for 13 speakers with 52 triples was chosen [40]. This incomplete design reduces the number of triples considerably, as a full design requires $\frac{n!}{3!\,(n-3)!}$ triples, which would result in 286 instead of 52 triples for 13 speakers. The validation of several incomplete designs with two different sets of English words shows sufficient reliability in [40], provided the number of word pairs occurs more than once (λ>1).

With the incomplete design applied in this paper, each speaker occurred 12 times, and every pairing of speakers appeared twice ($\lambda = 2$), each with a different third speaker within the triple. The sentence content was identical for each stimulus, but varied between triples, in order to avoid boredom, so 12 different stimuli were required for each speaker. One complete session took about one hour, with typically 50–55 min. for the main task.

The (de)briefing took about two minutes. The same procedure was applied for each of the three sets of speakers and listeners. Each set was grouped for gender and language, in order to reduce i.e. cross-gender effects: male German, female German and male Australian English speakers.

4. Material and Participants

For all three experiments, 13 speakers were selected from available databases of read speech, providing two sentences for each speaker. 15 same-sex listeners were recruited for each trial. The sentence selection was dominated by finding a set for each trial exhibiting no obvious repairs or pronunciation errors while still providing the same sentences within each triple for the combinations required by the incomplete design. This drastically restricted the available speakers.

For the German data, read utterances from the Phondat 1 database are used [41]. The Phondat 1 utterances are short German sentences with a length of one intonation phrase and everyday semantic content such as "A farmer is working on his field" (translation by the authors). From a pool of 406 sentences 13 male and 13 female speakers were selected. Speakers' information is not available, but the male listeners were aged 22–46 (M = 29.8, SD = 7.66) and the female listeners 21–36 (M = 27.3, SD = 4.25). No listener had a background in psychology, phonetics, or the like. They were paid for their contribution and were given AKG K-601 headphones for playback.

For the Australian experiment, data from the AusTalk database was used [42]. From this diversified database, read sentences were chosen. From 59 sentences provided, 54 sentences were selected. Although the sentence content implies more bias than the German sentences (e.g. "Troy flicks through a yuppie magazine when he got the chance."), all recordings sound quite matter-of-fact to the authors. As the listening participants were to be recruited from the Sydney region, male speakers from one of the three AusTalk Sydney recordings sites were selected. There were 15 male listeners (aged 18–44, M = 26.7, SD = 6.39) who all fulfilled the requirements defined earlier for the AusTalk speaker selection, namely having attended school only in Australia until the age of 18, and thus being Australian English speakers but not necessarily born in Australia, in order to represent the heterogeneous background of Australian people. All received either credit points or were paid for their contribution and used Sennheiser HD 650.

All utterances had been normalized in level to prevent loudness adjustments after training, but no further manipulation was conducted, as a holistic perception of voice and speaking style should not be inhibited. The sequences of triples were randomized in order to create individual playlists for each participant. Thus, comparison of the pairs chosen from each triple can be conducted to assess reliability. One additional triple, based on one additional sentence and three speakers who were not part of the actual test, was used for training.

5. Results

As an indicator of consistency between the participants' choices, the percentage of the mode was chosen, i.e. how many of the participants chose the most frequently selected pair for each triple. It is on average 62% (40%–93%) for German males and 62% (33%–100%) for German females, compared to a chance level of 33%. For the Australian men it is 61% (40%–93%). According to Fleiss' kappa value, there is "slight agreement" (κ = .201) for the German male group, "fair agreement" (κ = .225) for the German female group, and "slight agreement" (κ = .194) for the Australian men.

5.1. Dissimilarity Ratings

Data was aggregated for each speaker to abstract from sentence differences in order to concentrate on speaker dif-

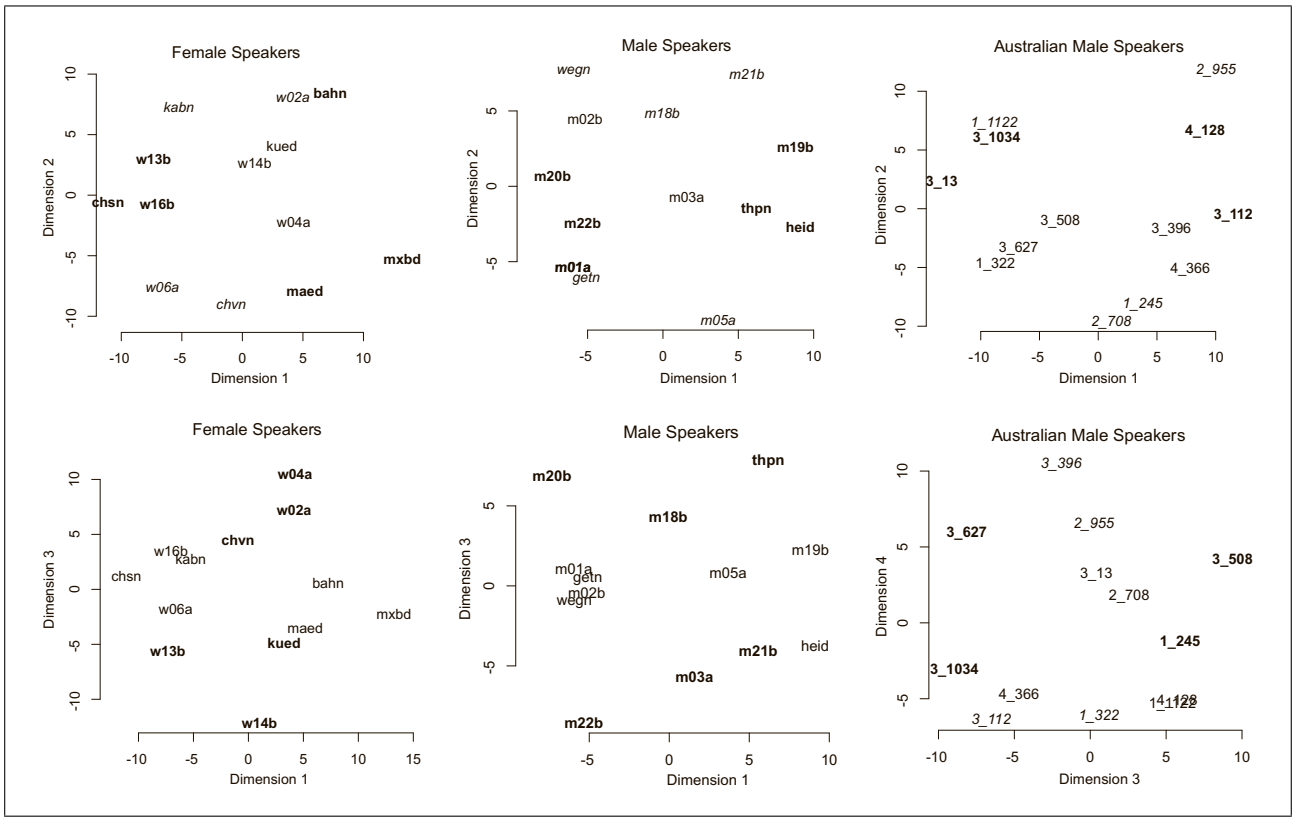

Figure 2. Results of the MDS solutions based on dissimilarity frequencies. Labels identify individual speakers from the databases. These speaker identifiers are bold, italic, or both if selected as extreme representatives for naming the respective dimension.

ferences. The distance between speakers for each of the three experiments was calculated by using the frequencies of dissimilar pairs. As the design did not include all possible combinations, applying non-metrical MDS on these distances was considered more appropriate than assuming Euclidean distances (executing monoMDS with the metaMDS R-package in order to find a stable solution). The results are presented for each data set.

Scree plots of the stress-1 values are evaluated to determine the number of dimensions [43]. As the performance values seem to converge and drop below a stress of .20 (stress for males: .124; stress for females: .128), as solution with three dimensions was chosen for both German sets. The Australian data is different, as there is one pronounced "elbow" favouring four dimensions (stress for Australian men: .050). Please see Figure 2 for the distances depicted for the first two dimensions separating the speakers for each data set.

5.2. Analysis of Individual Labels

The labels provided by the listeners were inspected for each dimension (translations of German labels given were translated by the authors). The main observation is that listeners did not restrict themselves to labels describing or valuing vocal features. Despite the instructions, they also inferred speaker states and speaker traits related to emotions or personality. This is in line with the behaviour reported in [11] and with theories explaining attribution of personality based on (and in this case applied to verbal) first impressions [44]. Given our previous experience obtained in the pre-test, we instructed the participants to avoid labels referring to emotion, attitude, or age, and to use only labels for stress if a speaker sounded very unnatural, as sentence stress differences may occur due to the reading situation of the recordings. However, this did not prevent a number of labels related to stress or the speaker states/traits.

In order to identify the dimensions found, the most extreme speakers were chosen: six for each dimension of the German data, and four for each dimension of the Australian data. In the latter case, choosing six speakers would have resulted in using far too many speaker pairs. From these six/four speakers, labels for all possible nine/four pairs were selected to represent maximum differences on this particular dimension, and as far as possible simultaneously exhibiting only minor differences in the other dimensions (Figure 2). Please recall that the labels associated to these pairings of extreme speakers are actually describing triples, as the assigned labels also refer to a third stimulus rated as similar to one of these two extreme stimuli. Therefore, actually all stimuli are included in this qualitative analysis, while labels for the most extreme ones are taken into account more frequently. There are 15 listeners multiplied with nine/four speaker pairs with two occurrences of each pair ($\lambda = 2$), resulting in a maximum of 270/120 labels per dimension. As you can see in Tables II–IV, the difference between the numbers of labels for the

German and Australian data is actually not that strong, as labels were disregarded if the selected similarity included two speakers from opposite ends of the dimension. This is often due to the third stimulus belonging to an extreme speaker as well.

The perceptual dimensions for the German women are named *tension*, *positive timbre*, and *maturity*, and those for the German men *calmness, factual*, and *naturalness*. The dimensions for the male Australian speakers are called *low pitch, remarkable timbre&voice*, and *emotion*, with an unnamed fourth dimension. The naming has been conducted with caution, and therefore the relationship of Unnamed to proficiency is not reflected due to sparse evidence. Likewise, all similarities between dimensions are presented in the Discussion, instead of overstressing similarity by choosing identical names. All names refer to the positive end of the scale.

To provide a systematic approach with a qualitative analysis of the labels, label pairs from these extreme pairs were clustered manually and categorized. Only a few labels lacking descriptive power were excluded (e.g. "similar vs. different pitch" instead of "high vs. low pitch"). The categories used are the ones from Table I. In cases of multi-category labels (e.g. "low–bright", "fluent–slow", "articulate–fast", "monotone–many accents"), the label describing the similarity was used for classifying, except for the case of it being a speaker trait or speaker state. In such a case, the non-speaker-related category was chosen (e.g. "less confident–more emphasis" → Accentuation).

As we encountered several instances of labels associated with stress ("strongly stressed–not stressed"), as well as with regularity of stress ("evenly stressed–unevenly stressed", "even meter–disjoint meter"), but also those concerned with fluency ("fluent–stagnant", "bumpy–smooth", "fluent–disjoint", "pause–no pause"), Accentuation was included and the categories Rhythm and Fluency were separated, although both seem to be somehow related (e.g. "stagnant"). And despite the rigorous and labour-intensive speaker selection process, a few labels refer to regional background (Accent). Tables II–IV present the number of antonyms mapped to each category, with the highest values marked in bold. Remarkable is the high number of labels used by the German participants related to Accentuation, Tempo, and Pronunciation compared to the Australian listeners. Labels for Vocal Effort, Rhythm, and Voice Quality, in contrast, are only mentioned infrequently. The high number of labels for speaker traits and speaker states were particularly helpful in interpreting the distribution of labels, e.g. the dimensions *maturity* and *naturalness*, as explained in the following dimension descriptions. The last line of each table reflects the positive or negative valence ratings of the labels, which are presented in the next section.

When interpreting the numbers in each category, consistency was also taken into account. For example, the labels "fast" and "slow" are consistently used, e.g., for the dimension *calmness*. However, they are not consistently used separating the extreme speakers for *naturalness* (i.e. "natural" and "unnatural" speakers are labelled as fast as well as slow), and are therefore not part of the character

Table II. Number of labels in each category by dimension (male Germans). Highest values in bold.

Categories	*calmness*	*factual*	*naturalness*
Pitch	08	15	08
Intonation	11	12	10
Accentuation	**30**	**32**	22
Rhythm	02	01	00
Fluency	10	14	10
Tempo	**58**	16	**48**
Pronunciation	**34**	12	25
Timbre	**32**	19	16
Voice quality	00	05	01
Vocal effort	00	07	02
Traits/States	**33**	**30**	**35**
Accent	01	00	03
Valence (%)	22	03	22

Table III. Number of labels in each category by dimension (female Germans). Highest values in bold.

Categories	*tension*	*positive timbre*	*maturity*
Pitch	22	22	**19**
Intonation	13	05	10
Accentuation	16	13	14
Rhythm	00	00	05
Fluency	17	05	**22**
Tempo	20	21	**21**
Pronunciation	15	27	15
Timbre	**28**	**34**	**22**
Voice quality	11	07	08
Vocal effort	07	03	03
Traits/States	**27**	25	**33**
Accent	**32**	**18**	00
Valence (%)	-42	22	30

Table IV. Number of labels in each category by dimension (male Australians). Highest values in bold. t&v: timbre&voice.

Categories	*low pitch*	*t&v*	*emotion*	–
Pitch	**23**	**18**	**12**	10
Intonation	03	05	02	04
Accentuation	02	03	00	01
Rhythm	02	00	00	02
Fluency	08	03	04	11
Tempo	11	06	**16**	07
Pronunciation	01	03	06	01
Timbre	**23**	**26**	**18**	**31**
Voice quality	07	06	04	06
Vocal effort	03	01	04	02
Traits/States	**23**	**16**	**16**	**24**
Accent	00	03	03	01
Valence (%)	10	-01	03	28

and not taken into account for the naming of these this dimension.

For each triple labelled the participants could also indicate if the label for the two similar stimuli or the third stimulus is considered as positive. To analyze these valence ratings the same 9, or 4, respectively, pairings of extreme stimuli are chosen for each dimension as for the naming of the dimension. For each possible pair between the four stimuli with the most positive dimension scores and the four stimuli with the most negative scores the ratings (either +1, 0, -1 from the viewpoint of the positive dimension scores) are summed up (please refer to Figure 2 for the selected stimuli). Therefore all stimuli are included in this analysis, however, ratings for the most extreme ones are more frequently taken into account. This sum reflects the difference between all positive and all negative evaluations and is given as percentage to all decisions for these 9 or 4 pairs, and subsequently is called valence (Tables II–IV).

The two German data sets reveal similar positive/negative evaluations: Over 80% (84–94) of the similarities taken into account are non-neutral compared to 61%–68% for the Australian data set. In consequence, valence is also higher for the German dimensions. *Calmness* and *naturalness* are considered positive, but *factual* is neutral. For female Germans, *positive timbre* and *maturity* have more positive evaluations, but *tension* is negative.

For the Australian speakers, 28% of all similarity ratings of the 4 pairs favour the speakers with high values on the proficiency-related unnamed dimension. The other dimensions exhibit low or non-existent valence differences.

6. Discussion

Qualitative methods are currently not as commonly used in voice analysis, as the required systematization process is inherently laborious and could easily take two more pages to describe. In the case of the Repertory Grid, such qualitative parts of the data are used to interpret the dissimilarities. This method does not generate a huge number of data points, but is considered necessary at this stage of research in order to elicit descriptors from the subjects. This objective has been reached. As a major result, meaningful perceptual dimensions have been found. Interpreting the labels provided by listeners failed only for one dimensi on (the 4th for the Australian data), while the valence rating supports the general findings. Here is a selection of labels (including German originals in brackets) for each dimension:

German males:

- calmness vs. activity (Gelassenheit–Aktivität)
  - slow, precise, stressed, clear (langsam, präzise, betont, klar)
  - hectic, unsettled, fast (hektisch, unruhig, schnell)
- factual vs. emotional (sachlich–emotional)
  - non emphasized, monotonous, clear, precise, fluent, neutral, unexcited (unbetont, monotone/flach, klar, deutlich, flüssig, neutral, unaufgeregt)
  - feigned, over-emphasized, over-articulate, halting, emotional, excited (affektiert, über-betont, überdeutlich, stockend, emotional, aufgeregt)
- naturalness vs. unnaturalness (Natürlichkeit–Unnatürlichkeit)
  - slower, authentic, natural, well stressed, soft (langsamer, authentisch, natürlich, gut/richtig betont, weich)
  - unbelievable, fast, unnatural, badly stressed (unglaubwürdig, schnell, unnatürlich, schlecht betont)

German females:

- tension vs. relaxation (Anspannung–Entspannung)
  - high, clear, bright, shrill, odd, excited, annoying (hoch, klar, hell, schrill, kalt, schräg, aufgeregt, nervig)
  - low, dark, relaxed, calm (tief, dunkel, entspannt, ruhig)
- positive vs. negative Timbre (positive vs. negative Klangfarbe)
  - clear, bright, sonorous, warm (klar, hell, klangvoll, warm)
  - cold, dull, dark, coarse (kalt, dumpf, dunkel, rau)
- Maturity vs. Immaturity (Reife–Unreife)
  - vivacious, grown-up, mature, professional, serious, benevolent, fluent, clear (vital, erwachsen, reif, professionell, ernst, wohlwollend, flüssig, klar)
  - peculiar, childish, unprofessional, bumpy (auffällig, kindlich, unprofessionell, holprig)

Australian males:

- Low vs. high pitch
  - low, dark (tief, dunkel)
  - cold, sharp, high (kalt, scharf, hoch)
- unremarkable vs. remarkable timbre&voice
  - smooth, clear, soft, clean, normal
  - hollow, heavy, rough, unclear, nasal, squeaking, scratchy
- emotion (low vs. high activity)
  - less energy, clear pronunciation, slow, sleepy, sad, calm, boring
  - energetic, fast, interested
- unnamed
  - warm, heavy, deep
  - even metre, more dynamic, motivated, serious, emphasized, interesting

Low pitch exhibits a strong descriptive character and refers to a basic attribute of voices. It might be related to only a few acoustic features like fundamental frequency and spectral centre of gravity. The other two relatively simple dimensions are *calmness*, which is predominantly related to Tempo and a proper Accentuation, and *factual*, for which normal versus overemphasized Accentuation are separated. However, all the other dimensions discussed here have to be considered as quite complex.

As finding reliable acoustic correlates for such concepts like tension [45] is still ongoing research, no acoustic analysis of these few and phonetically diverse stimuli will be

presented here. Instead, a questionnaire to describe speakers from a layman's perspective will be revised include the identified dimensions in order to assess a proper amount of data for acoustic analysis. *Positive timbre* and *remarkable timbre&voice*, for example, comprise many different aspects such as "nasal", "rough", "warm" etc. *Calmness, tension, maturity, naturalness*, and *emotion* are described by labels from various categories, with *tension* and *emotion* apparently related to physiological speaker states.

There are similarities between the data sets of the two languages. *Calmness* (as opposed to activity), *tension* and also *emotion* are related, and *naturalness* and *maturity* both cover the aspect of proficiency. Also, some kind of timbre occurs in two data sets (German females and Australian males). Despite the unique naming of the dimensions to reflect the connotations perceived, such similarities provide additional evidence of perceptual dimensions related to arousal/tension, timbre, and speaking proficiency.

Comparing these results with the pre-test conducted for other speakers, dimensions related to *calmness* and *naturalness* have been found as well.

Additionally, there are commonalities with most categories already found in the related studies presented in Table I. These are the categories of Pitch (*low pitch*), Intonation (*factual*), Timbre (*timbre, timbre&voice*), others (affectation) (*emotion*), as well as Tempo as expression of *calmness*. Other dimensions have to be considered as too complex for a one-to-one mapping to the categories: *naturalness*, *proficiency*, and especially *tension*.

We want to point out that this complex character of the dimensions and the frequent occurrence of speaker states and speaker traits cannot be considered as artefacts of inconsistent labelling by different participants. The label pairs provided give insight into the percepts of the listeners, indicating already complex and multifaceted descriptors which are not true antonyms anymore. Consider the examples of "low–bright", "fluent–slow", and "articulate–fast" for the simplest cases of a relation between Pitch and Timbre, Fluency and Tempo, or Pronunciation and Tempo. Other examples are even more interesting: "low toned–motivated", "more emphasis–less confident", "emotional–flat", "light–aggressive", "vibrating–smooth", "relaxed-high". Because of this, labels cannot always be divided into descriptive labels ("high", "fast" , "shrill") and attributional labels ("awake", "friendly", "young"). In spite of contrary instructions, there are frequent occurrences of speaker states and speaker traits. Subsequently, many dimensions names are also related to traits and states (*calmness, tension, emotion, maturity*, maybe also *naturalness*). Only *factual*, *low pitch*, and the two timbre related dimensions refer to the perception of voice and speaking style.

As there is no definitely unique dimension for the German female speakers, no obvious gender effect is visible. This may of course change when applying cross-gender ratings or repeating this experiment. However, the two male studies evidently differ in the dimensions. A cross cultural experiment relying on identical stimuli might reveal, if for example the difference in importance of pronunciation is due to the set of speakers, listeners, or if it represents an actual cultural difference.

7. Conclusion

The procedure for finding speaker differences expressed in distances is attractive, as it does not bias participants due to the item labels. However, the process of naming dimensions with this procedure is much more challenging than using a questionnaire, as typically done [9, 24].

The non-expert listeners could apparently not refrain from making state or trait attributions instead of just describing voice and speaking style. Such a behaviour is not reflected in questionnaires with fixed scales. The consequence is a multifaceted character of the dimensions expressed by varying categories of labels, compared to the factual and descriptive dimensions found in the literature. Still, no absolutely new aspects are found. Instead, there is additional evidence for some of the dimensions, actually for most of the categories of dimensions as depicted in Table I. Nevertheless, the dimensions which were found in this study are more complex and abstract, and relate strongly to speaker states and speaker traits. For example, the categories Intonation, Rhythm, Voice Quality, and Vocal Effort are reflected in the dimensions found.

Therefore, the status of perceptual descriptors from experts and non-experts has to be considered as different, due to the top-down influence of person attributions evident in the non-experts. For the aim of precisely describing voice and articulation, only professionals should be recruited. One example application is the identification of acoustic correlates of voice qualities and timbre. Results from non-experts describing speakers' voices and speaking styles on questionnaires specialized on phonetic details, however, should be treated with care. For such non-experts, a more abstract questionnaire should be applied instead, which covers aspects found in our data, such as tension, proficiency, or naturalness. If validly assessed, such concepts represent a stage in the cognitive processing are therefore important for studying the formation of voice-based attributions in ordinary people, as illustrated, e.g., with the Brunswikian lens model [1].

For its development, labels of the scales have been included and/or adjusted based on the labels used here by the participants. More importantly, a subsequently revised questionnaire [46] includes attributions regarding the dimensions found here; for example, "active-passive", "emotional-neutral", "interested-indifferent", "mature-childish", and "feminine-masculine" particularly addressing speaker characterization, but "tensed-relaxed", "professional-ordinary", "natural-unnatural", and "pleasant timbre-unpleasant timbre" for voice descriptions. A preliminary validation resulted in the descriptive dimensions softness, fluency, activity, precision, darkness, and tempo [47]. With this questionnaire, the stability of the dimensions found here can be examined as follow-up work,

further validating the current state in perceptual dimensions.

Acknowledgement

This work was financially supported by the German Research Foundation, DFG (grant WE 5050/1-1).

References

[1] K. Scherer: Methods of research on vocal communication: paradigms and parameters. – In: Handbook of methods in nonverbal behavior. K. Scherer, P. Ekman (eds.). Cambridge University Press, Cambridge, 1982, 136–198.

[2] K. Scherer: Vocal communication of emotion: A review of research paradigms. Speech Communication **40** (2003) 227–256.

[3] M. D. Back, S. C. Schmukle, B. Egloff: A closer look at first sight: Social relations lens model analysis of personality and interpersonal attraction at zero acquaintance. European Journal of Personality **25** (2011) 225–238.

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

[5] M. Hirano: Clinical examination of voice. Springer, New York, 1981.

[6] T. Bänziger, S. Patel, K. R. Scherer: The role of perceived voice and speech characteristics in vocal emotion of communication. Journal of Nonverbal Behavior **38** (2014) 31–52.

[7] L. Boves: The phonetic basis of perceptual ratings of running speech. Foris Publications, Dordrecht, 1984.

[8] B. Ketzmerick: Zur auditiven und apparativen Charakterisierung von Stimmen. TUDpress, Dresden, 2007, (Studientexte zur Sprachkommunikation).

[9] K. Scherer: Voice quality analysis of American and German speakers. Journal of Psycholinguistie Research **3** (1974) 281–298.

[10] B. Weiss, F. Burkhardt: Voice attributes affecting likability perception. Proc. Interspeech, 2010, 1934–1937.

[11] J. Laver: Labels for voices. Journal of the International Phonetic Association **4** (1974) 62–75.

[12] N. Schwarz, F. Strack, H. P. Mai: Assimilation and contrast effects in part-whole question sequences: A conversational logic analysis. Public Opinion Quarterly **55** (1991) 3–23.

[13] P. Susini, G. Lemaitre, S. McAdams: Psychological measurement for sound description and evaluation. – In: Measurement with Persons: Theory, Methods, and Implementation Areas. B. Berglund, G. Rossi, J. Townsend, L. Pendrill (eds.). Psychology Press, New York, 2012, 227–253.

[14] J. Kreiman, B. R. Gerratt: The perceptual structure of pathologic voice quality. Journal of the Acoustical Society of America **100** (1996) 1787–1795.

[15] F. Nolan, K. McDougall, T. Hudson: Some acoustic correlates of perceived (dis)similarity between same-accent voices. Proc. of the 17th International Congress of Phonetic Sciences, 2011, 1506–1509.

[16] M. Wältermann: Dimension-based quality modeling of transmitted speech. Springer, Berlin, 2013.

[17] B. Schuller, A. Batliner: Computational paralinguistics: Emotion, affect and personality in speech and language processing. Wiley, 2013.

[18] J. McWhorter: Lexicon valley podcast no. 24. https://soundcloud.com/panoply/lexicon-valley-24-get-your#t=4:31, Mon, 31 Dec 2012.

[19] R. Anderson, C. A. Klofstad, W. J. Mayew, M. Venkatachalam: Vocal fry may undermine the success of young women in the labor market. PLoS ONE **9** (2014) e97506.

[20] I. Yuasa: Creaky voice: A new feminine voice quality for young urban-oriented upwardly mobile american women? American Speech **85** (2010) 315–337.

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

[22] T. Murry, S. Singh: Multidimensional analysis of male and female voices. Journal of the Acoustical Society of America **68** (1980) 1294–1300.

[23] T. Murry, S. Singh, S. Sargent: Multidimensional classification of abnormal voice qualities. Journal of the Acoustic Society of America **61** (1977) 1630–1635.

[24] W. D. Voiers: Perceptual bases of speaker identity. Journal of the Acoustical Society of America **36** (1964) 1065–1073.

[25] I. V. Bele: Dimensionality in voice quality. Journal of Voice **21** (2007) 257–272.

[26] B. Weiss, S. Möller: Wahrnehmungsdimensionen von Stimme und Sprechweise. Proc. Elektronische Sprachsignalverarbeitung (ESSV), 2011, 261–268.

[27] W. Fagel, L. V. Herpt: Analysis of the perceptual qualities of Dutch speakers' voice and pronunciation. Speech Communication **1** (1983) 315–326.

[28] J. Kreiman, B. R. Gerratt, G. B. Kempster, A. Erman, G. S. Berke: Perceptual evaluation of voice quality: Review, tutorial, and a framework for further research. Journal of Speech and Hearing Research **36** (1993) 21–40.

[29] O. Baumann, P. Belin: Perceptual scaling of voice identity: common dimensions for different vowels and speakers. Psychological Research **74** (2010) 110–120.

[30] J. Kreiman, G. Papcun: Comparing discrimination and recognition of unfamiliar voices. Speech Communication **10** (1991) 265–275.

[31] S. Singh, T. Murry: Multidimensional classification of normal voice qualities. Journal of the Acoustical Society of America **64** (1978) 81–87.

[32] K. Scherer, H. Giles: Social markers in speech. Cambridge University Press, 1979.

[33] E. Strangert, J. Gustafson: What makes a good speaker? subjective ratings, acoustic measurements and perceptual evaluation. Proc. Interspeech, 2008, 1688–1691.

[34] I. R. Titze: Acoustic interpretation of resonant voice. Journal of Voice **15** (2001) 519–528.

[35] B. Weiss, F. Burkhardt, M. Geier: Towards perceptual dimensions of speakers' voices: Eliciting individual descriptions. Proc. Workshop on Affective Social Speech Signals, 2013, 1–5.

[36] G. Kelly: The psychology of personal constructs. Norton, New York, 1955.

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

[38] F. Fransella, R. Bell, D. Bannister: A manual for repertory grid technique. 2 ed. Wiley, Chichester, 2004.

[39] M. Hassenzahl, R. Wessler: Capturing design space from a user perspective: the repertory grid technique revisited. International Journal of Human?Computer Interaction **12** (2000) 441–459.

[40] M. L. Burton, S. B. Nerlove: Balanced designs for triads tests: two examples from English. Social Science Research **5** (1976) 247–267.

[41] Bayerisches Archiv für Sprachsignale: PhonDat 1. 1995.

[42] D. Estival, S. Cassidy, F. Cox, D. Burnham: Austalk: an audio-visual corpus of australian english. Proc. Language Resources Evaluation Conference (LREC), 2014, 3105–3109.

[43] I. Borg, P. Groenen: Modern multidimensional scaling. 2nd ed. Springer, New York, 2005, (Springer Series in Statistics).

[44] H. H. Kelley: Causal schemata and the attribution process. – In: Attribution: Perceiving the causes of behavior. E. E. Jones, D. Kanouse, H. Kelley, R. Nisbett, S. Valins, B. Weiner (eds.). General Learning Press, New York, 1972, 1–26.

[45] D. Bone, C.-C. Lee, S. Narayanan: Robust unsupervised arousal rating: A rule-based framework with knowledge-inspired vocal features. IEEE Trans Affect Comput. **5** (2014) 201–213.

[46] L. Fernandez Gallardo, B. Weiss: The Nautilus speaker characterization corpus: Speech recordings and labels of speaker characteristics and voice descriptions. submitted to LREC, 2018.

[47] B. Weiss: Voice descriptions by non-experts: Validation of a questionnaire. Proc. Phonetics & Phonology, 2016, 228–231.