Qual User Exp (2017) 2:2
https://doi.org/10.1007/s41233-017-0005-6

RESEARCH ARTICLE

Latent factor analysis for synthesized speech quality-of-experience assessment

Rishabh Gupta[1] · **Tiago H. Falk**[1]

Received: 9 March 2016 / Published online: 6 February 2017

Abstract Text-to-speech (TTS) systems are evolving and making way into numerous commercial systems, such as smartphones and assistive technologies. Notwithstanding, their user perceived quality-of-experience (QoE) is still low compared to natural speech, with distortions arising across numerous perceptual dimensions, such as voice pleasantness, comprehension, and appropriateness of intonation, to name a few. Unfortunately, the effects of such perceptual dimensions on overall perceived QoE is still unknown, particularly across listeners of different genders, thus making it difficult for TTS developers to further improve system quality. To overcome this limitation, this study makes use of exploratory factor analysis (EFA), confirmatory factor analysis (CFA), and model invariance tests to shed light on factors responsible for QoE perception across natural and synthesized speech, as well as male and female listeners. Experimental EFA/CFA results on a publicly available database of commercial TTS systems showed the emergence of two key perceptual dimensions responsible for TTS QoE, namely 'listening pleasure' and 'prosody'. Model invariance tests validated the reliability of the model across male and female listeners, as well as across natural and synthetic voices.

Keywords Confirmatory factor analysis · Exploratory factor analysis · QoE · TTS · Model invariance

✉ Tiago H. Falk
falk@emt.inrs.ca

[1] INRS-EMT, University of Quebec, Montreal, Canada

Introduction

Modern day text-to-speech (TTS) systems have shown tremendous progress since the bandpass-based Voder developed by Dudley in 1939. The quality of modern TTS systems has reached a level that allows leveraging synthetic speech in everyday applications, such as audiobooks, smartphones, computers, assistive technologies and global positioning systems, to name a few (Hinterleitner et al. 2012). TTS systems have also gained great popularity in the domain of personal digital assistants (PDAs), such as Apple's Siri, Google Now from Google, and Cortana from Microsoft. The success of these emerging TTS applications and systems, however, requires systems to output synthesized speech of high quality, thus delivering an optimal quality-of-experience (QoE) to the user (Hinterleitner et al. 2014).

Back in the late 1970's, parametric TTS systems, such as the formant synthesizer, became popular as they were the first systems that could produce intelligible synthetic speech (Klatt 1980). The generated speech from such systems, however, sounded artificial and robotic (Hinterleitner et al. 2014). Corpus-based synthesizers were later introduced and made use of concatenated diphone units; such systems, however, suffered from sonic glitches that occurred at the conjunction of two units (Hinterleitner et al. 2014). Unit-selection based TTS systems, in turn, appeared in the mid-90's (e.g., Black and Taylor 1994) and relied on the selection of units from a large database of prerecorded speech while minimizing a cost function. While such systems sound very natural, poor clarity and intelligibility of short segments has been reported. The latest developments in speech synthesis are the hidden Markov model (HHM) based systems (Tokuda et al. 2002) which are trained on excitation and spectral parameters of human

speech. The naturalness of HMM-synthesizers has been shown to be subpar compared to unit-selection synthesizers. Notwithstanding, they do not suffer from the prosodic glitches seen with concatenation-based systems (Hinterleitner et al. 2014).

Speech impairments generated by existing TTS systems degrade the perceived quality along different perceptual dimensions or constructs, thus highlighting the multidimensional nature of synthetic speech quality (Hinterleitner et al. 2012). Previous research using subjective listening tests have tried to analyze these underlying perceptual dimensions. In Kraft and Portele (1995), five different TTS systems were analyzed using exploratory factor analysis (EFA), thus revealing two main perceptual dimensions (or factors) related to TTS quality. The first factor measured the prosodic and long-term attributes, whereas, the second factor represented segmental attributes. These findings, however, did not include unit-selection or HMM based synthesizers. In Mayo et al. (2005), in turn, a multidimensional scaling (MDS) analysis revealed three perceptual dimensions for unit-selection based TTS quality, namely, prosody, appropriateness, and number of selected units. More recently, EFA was performed using speech stimuli from a wide variety of TTS systems, consisting of both male and female voices (Hinterleitner et al. 2011a). The listeners scored their perceived quality across several indicators, including naturalness of accentuation, pleasantness, bumpiness, noisiness, intelligibility, and rhythmicity, amongst others. EFA analysis reduced these indicators into three relevant perceptual dimensions, namely, naturalness, disturbances and temporal distortions (Hinterleitner et al. 2011a). A similar study with audiobooks revealed two major perceptual dimensions: listening pleasure and prosody (Hinterleitner et al. 2011b). Lastly, MDS analysis showed three major perceptual dimensions, namely naturalness, temporal distortions, and calmness with these dimensions shown to be correlated with voice pleasantness and intelligibility, rhythm and fluency, and speed, respectively (Hinterleitner et al. 2012).

As can be seen, MDS and EFA have been widely used to extract (latent) perceptual dimensions involved in the evaluation of perceived TTS quality, thus allowing for psychometric models to be developed linking latent factors to perceived quality indicators. The above-mentioned models, however, have two major limitations. First, the goodness-of-fit (GOF) of the developed psychometric models was not measured, thus casting doubt on their generability. Second, invariance of the model across different groups, such as listener gender, was not explored. Listener gender has been shown in the past to be a potential influential factor in TTS assessment (Mullennix et al. 2003), thus further work is needed. Here, we overcome these limitations by leveraging the use of confirmatory factor analysis (CFA) in addition to EFA. CFA extracts the goodness-of-fit of the obtained models, thus measures the reliability and validity of the developed psychometric model (Viswanathan and Viswanathan 2005). Moreover, CFA is also used to establish model equivalence (or invariance) across listener gender. As an additional step, we also performed model equivalence between natural and synthesized speech generated from current state-of-the-art commercial personal digital assistants (PDAs). To the best of authors' knowledge, such comprehensive psychometric analysis of natural and synthesized speech quality has not been reported previously.

The remainder of this paper is organized as follows: Sect. 2 describes the experimental design used for data collection. Section 3, in turn, describes the EFA, CFA and model invariance analyses. Sections 4 and 5 show the experimental results and discussion, respectively. Lastly, conclusions are drawn in Sect. 6.

Experimental setup

This section details the participants, speech stimuli, rating dimensions, and experimental protocol used for data collection. Data was collected over two sessions. Data from the first session (pilot) was used for EFA and from the second (main) for CFA and model invariance testing/validation. Data from the second session has been made publicly available (Gupta et al. 2015).

Participants

A total of 28 participants were recruited for the study, six of which participated in session one (pilot) and 21 in session two (main). All participants were fluent in English. For session one, two were female and the participant average age was 31.16 ($\pm$8.18). For session two, (eight females), the average age was 23.8 ($\pm$4.35). None of the participants reported having any hearing or neuro-physiological disorders. The study protocol was approved by the INRS Research Ethics Office and participants consented to participate in the studies. Participants were compensated monetarily for their time.

Speech stimuli

Speech stimuli used are listed in Table 1, along with number of male/female voice recordings and sentence duration. Stimuli consisted of four natural voices and seven synthesized voices obtained from commercially available systems namely, Microsoft, Apple, Mary TTS Unit selection and HMM, vozMe, Google and Samsung. Tested systems cover a range of different concatenative, unit

Table 1 Description of the stimuli used for the listening tests

Type	System	Sentence group	Male sets	Female sets	Duration range (s)
Natural	1	A	0	4	17–19
	2	A	0	4	18–23
	3	A	0	4	17–19
	4	B	0	4	13–14
Synthesized	5	A	0	4	19–24
	6	A	0	4	17–22
	7	A	2	2	17–20
	8	A	2	2	18–25
	9	A	2	2	17–22
	10	A	2	2	17–21
	11	A	2	2	13–17

selection and HMM-based systems. A non-identifying code is provided for the four natural voices and seven TTS systems in Table 1. Speech samples were generated from two sentence groups (A and B), each comprising of four sentences. The content of sentence groups A and B differed from each other slightly. Also, the sentences in group B were slightly shorter as compared to the sentences in group A. Thus, the total number of stimuli used in this study were 44 (natural voices: 4 + synthesized voices: 7 = 11 voices × 4 sets of sentences = 44 stimuli). The speech stimuli also consisted of both male and female voiced sets of sentences for five of the seven synthesized voices. The speech stimuli were presented to listeners via headphones at a sampling rate of 16 KHz and a bitrate of 256 kbps.

Subjective rating dimensions

In our studies, we presented listeners with 12 subjective rating scales to gauge their perception of quality and quality-of-experience (QoE). To this end, typical quality-related ratings were used, such as those in Hinterleitner et al. (2011b), as well as users affective state ratings, which are useful for QoE measurement (Brunnström et al. 2013). All the items were scored on a continuous scale. The 12 ratings used are listed below and more details are listed in Table 2, including the abbreviations of each dimension used throughout the remainder of this paper. It should be noted that the affective dimensions of valence, arousal and dominance were measured using the self assessment manikins (SAM) (Morris 1995). SAM is a 9-point non-verbal pictorial assessment technique for affect measurement. While measuring valence, the first and last pictures represent negative and positive pleasantness, respectively. For arousal the first and last pictures represent unexcited and excited, respectively. Finally for dominance they represent not in control and in control nature of affect.

1. *Overall impression* This scale evaluated the overall quality of the system considering all the aspects.
2. *Voice pleasantness* This measured the degree of voice pleasantness.
3. *Speaking rate* This measure reflected the listener's reaction to the speed of delivery in a real situation.
4. *Acceptance* This scale measured whether the voice could be accepted as a Personal Digital Assistant or not.
5. *Intonation* This scale gauged whether the produced pitch curve fits to the sentence type.
6. *Naturalness* This scale measured the level of naturalness/unnaturalness of the voice.
7. *Listening effort* This captured the effort required to listen to a particular voice while listening to it for a longer duration of time.
8. *Comprehension problems* This scale measured the comprehension problems that might have arisen due to badly synthesized speech.
9. *Emotion* This item captured the variations of voice which reflected the atmosphere of the scene being described.
10. *Valence* This item captured the attractiveness (positiveness) or averseness (negativeness), of the voice, as experienced by the listener.
11. *Arousal* This item measured the level of mental alertness/excitation of the listener after listening to the voice.
12. *Dominance* This item measured the feeling of control over the situation after listening to the voice.

Experimental protocol

The experimental procedure was carried out in accordance with ITU-T P.85 recommendations (ITU-T 2016), with no secondary task. Participants were comfortably seated in front of the computer screen inside a sound proof room.

Table 2 Subjective dimensions used in the listening test along with their description and abbreviations used herein

Dimensions	Abbreviation	Recommendation	Description
Overall impression	MOS	ITUT-T P.85 (ITU-T 2016)	1-Bad,... 5-Excellent
Voice pleasantness	VP	ITUT-T P.85 (ITU-T 2016)	1-Very unpleasant,... 5-Very Pleasant
Speaking rate	SR	ITUT-T P.85 (ITU-T 2016)	1-Slow,... 5-Fast
Acceptance	Acc	ITUT-T P.85 (ITU-T 2016)	1-Strongly don't accept,... 5-Strongly accept
Intonation	Int	Hinterleitner et al. (2011b)	1-Melody did not fit sentence type,... 5-Melody fitted the sentence type
Naturalness	Nat	Hinterleitner et al. (2012)	1-Unnatural,... 5-Natural
Listening effort	LE	ITUT-T P.85 (ITU-T 2016)	1-Very exhausting,... 5-Very Easy
Comprehension problems	CP	ITUT-T P.85 (ITU-T 2016)	1-Never,... 5-All the time
Emotions	Emo	Hinterleitner et al. (2011b)	1-No expression of emotions,... 5-Authentic expression of emotions
Valence	Val	New	1-Negative,... 9-Positive
Arousal	Ar	New	1-Unexcited,... 9-Excited
Dominance	Dom	New	1-Not in control,... 9-In control

Insert earphones were placed comfortably inside the participants' ears to deliver the speech stimuli at their individual preferred volume levels. The experiment was then carried out in two phases: a familiarity phase and an experimental phase. In the familiarity phase, participants were presented with a sample speech file followed by the series of rating questions, thus illustrating the experiment procedure and giving them the opportunity to report any problem and/or concerns. In the experimental phase, participants were presented with a randomized speech stimuli, one sentence set (approximately 20 s long) at a time. Following each stimulus, participants were presented with a randomized series of rating questions on the screen wherein the participants scored the stimuli on the rating scales described in Table 2.

Factor analysis

Factor analysis is a multivariate statistical technique which was developed to test hypotheses regarding the correspondence between scores on observed variables (surface attributes), or indicators, and the hypothetical constructs (internal attributes), or latent factors, presumed to affect such scores (Kline 2013). The foundation of factor analysis is the assumption that the internal attributes exist. The internal attributes are the hypothetical constructs that can be used for understanding and accounting for the observed phenomenon. The internal attributes are more fundamental than surface attributes and can not be measured directly; however, their effects are reflected from the measures of surface attributes. The basic principle of factor analysis is that the internal attributes influence the surface attributes in a systematic manner, thus, measurements obtained from indicators are, at least in part, the result of the linear influence of the underlying latent factors (Tucker and MacCallum 2016).

Factor analysis has three major applications. First, it can be applied for the reduction of the number of indicators into a smaller set. Second, it can be used to establish the underlying dimensions between the indicators and the latent factors, thus generating or refining the theory. Finally, factor analysis provides construct validity evidence of the self-reporting scales (Thompson 2004; Tabachnick and Fidell 2001; Taherdoost et al. 2014). There are two discrete categories of factor analysis techniques: exploratory factor analysis (EFA) and confirmatory factor analysis (CFA). The EFA estimates unrestricted measurement models whereas, CFA analyses restricted measurement models (Kline 2013). Thus, for CFA the indicator-factor correspondence needs to be specified, whereas, for EFA there are no specific expectations regarding number or nature of underlying factors. The EFA and CFA techniques are further described in the subsections below.

Exploratory factor analysis

Exploratory factor analysis allows researchers to explore the main dimensions to generate a theory, or model from a relatively large set of indicators (Thompson 2004; Pett et al. 2003; Taherdoost et al. 2014). The EFA is particularly suitable for scale development and applied when the theoretical basis for specifying the numbers and patterns of common latent factors is unavailable (Taherdoost et al. 2014). The ultimate goal of EFA is to determine the number of latent factors that are required to explain the correlations between the indicators, thus, establishing the theory. The EFA is based on the common factor model that postulates that each indicator in a set of indicators is a linear function of one or more common factors and a

unique factor (Thurstone 1947). The common factors are the unobservable latent factors that influence more than one indicator in a set of indicators and are presumed to account for the correlations among the indicators. The unique factors are the latent variables that are assumed to influence only one indicator from a set of indicators and do not account for the correlations among the indicators. The objective of common factor model is to understand the structure of correlations among the indicators by estimating the relationship patterns between indicators and latent factors indexed by so-called *factor loadings* (Fabrigar et al. 1999). The goals of EFA for the current study were two-fold: (1) probe the validity of the factor structure obtained from Hinterleitner et al. (2011b), and (2) explore the measured affective "loadings," i.e. valence, arousal and dominance, on the obtained factors.

The EFA approach is sequential and linear, and involves many options, therefore, development of a protocol for analysis is imperative. There are several methodological issues associated with the EFA procedure, one of them being the indicator selection process. The indicator selection is an absolutely critical step as it determines the quality of the factor analysis (Fabrigar et al. 1999). Therefore, for the current study the indicator selection process involved selection of indicators based on P.85 recommendations (ITU-T 2016) and previous research (Hinterleitner et al. 2011a, b), that has helped in identifying important hypothetical constructs or latent factors associated with synthesized speech QoE.

A second methodological issue relates to the sufficiency of available data for EFA. The first consideration towards establishing the sufficiency of data is sample size. Various recommendations and opinions exist regarding the optimum sample size for EFA. For example Comrey and Lee (2013) suggested that a sample size of 50 is very poor, 100 is poor, 200 is fair, 300 is good, 500 is very good and 1000 is excellent. Moreover, MacCallum et al. (1999) illustrated that with commonalities greater than 0.6 and with each latent factor defined by several indicators (Henson and Roberts 2006), sample size can be relatively smaller. Other studies, in turn, have suggested that the nature of the data is what should determine the adequacy of the sample size (Fabrigar et al. 1999; MacCallum et al. 1999). Another recommendation towards establishing sample size adequacy is based on sample to variable ratio, denoted as *N:p* where *N* refers to the sample size and *p* refers to number of indicators. The rules of thumb for N:p values have ranged from 3:1 to 20:1 in the literature (e.g., see Costello and Osborne 2005).

An additional consideration towards establishing data sufficiency is the factorability of the correlation matrix. A factorable matrix consists of several sizeable correlations, therefore, the correlation matrix must be inspected for correlations above 0.30 for factor analysis to be meaningful (Tabachnick and Fidell 2001). Finally, the so-called Kaiser–Meyer–Olkin (KMO) measure (Kaiser 1970) and Bartlett's test of sphericity Bartlett (1950) have been proposed as measures of accurate sampling adequacy (Taherdoost et al. 2014). The KMO measure is indicative of the proportion of variance among the items that is common, thus suggesting an underlying latent factor. The KMO measure varies between 0 and 1, and values above 0.5 are typically considered to be adequate for EFA (Kim and Mueller 1978). The Bartlett's test of sphericity, on the other hand, tests the hypothesis that the correlation matrix is an identity matrix, suggesting that all variables are uncorrelated (Hair et al. 2009). If significance values are found lower than an alpha level of 0.05, the null hypothesis is rejected, thus suggesting that the correlation matrix is the identity matrix and that items are unrelated. In the current EFA study, the sample size (N) was 264, as 6 subjects scored 44 speech stimuli, and the number of indicators (p) used were 11 thus, leading to a $N{:}p$ ratio of 24:1. The KMO measure and Bartlett's test of sphericity are also used herein to establish sample adequacy.

A third methodological issue in performing EFA relates to the factor extraction method. There exists several factor extraction methods, such as principal component analysis (PCA), principle axis factoring (PAF), and maximum likelihood (ML) (Costello and Osborne 2005; Hair et al. 2009). The PCA based method computes factors without any regard to the underlying latent factors, whereas the PAF based method is used for the determination of the underlying latent factors related to the indicators (Taherdoost et al. 2014; Fabrigar et al. 1999). The maximum likelihood based method, in turn, is more suitable when the data is normally distributed and allows the computation of various goodness-of-fit measures for the model (Fabrigar et al. 1999). The PCA and PAF based methods are the most commonly used methods for EFA (Taherdoost et al. 2014). In the present study, PAF based factor extraction was used as it does not require the data to be normally distributed and is less likely to produce improper results compared to ML based methods (Fabrigar et al. 1999).

A fourth methodological issue involves choosing the factor retention method. The number of factors to be retained is an important consideration as under- or over-extraction of factors can result in substantial errors, thus affecting the efficiency and meaning of EFA (Taherdoost et al. 2014). There are various criteria for factor retention, such as Kaiser's criterion and Scree test. Kaiser's criterion recommends to retain all the latent factors that have eigenvalues greater than one, as this is the average size of eigenvalues in the full decomposition (Kaiser 1960). The Scree test, in turn, recommends to explore the graphical representation of the eigenvalues for discontinuities, as the

number of data points above the discontinuity represents the major factors (Hair et al. 2009). In this study, the number of factors to retain were determined using a combination of both Kaiser's criterion and the Scree test.

Another EFA related methodological issue involves the selection of the rotation method. The rotation of factors helps to produce simplified and interpretable results by maximizing high factor loadings and minimizing low factor loadings. There are two categories of rotation techniques namely, orthogonal and oblique rotation. Orthogonal rotation produces factors that are uncorrelated to each other, whereas oblique rotation results in factors that correlate to each other, thus leading to the production of correlated construct structures (Costello and Osborne 2005). There exists various methods for orthogonal and oblique rotation, such as varimax, quartimax and equamax for orthogonal rotation and quartimin and promax for oblique rotation (Mulaik 2009). In the current study, we used the promax oblique rotation method for EFA as it produces simplified factor structures while minimizing the cross-loadings (Hinterleitner et al. 2011b). The promax rotation begins with varimax rotation followed by raising the pattern coefficients to a higher power κ (Kappa), that forces near-zero coefficients to approach zero faster (Mulaik 2009). The κ value usually ranges between 1 and 4, and for the current study we used a κ value of 4, as in (Hinterleitner et al. 2011b).

Lastly, the final issue relates to the interpretation of the produced factor structure and naming the construct based on the factor loadings. This final step reflects the theoretical and conceptual intent and allows for better model interpretation (Hair et al. 2009). In order to meaningfully interpret a factor, at least 2 to 3 indicators must load onto it. The theoretical and conceptual interpretation of the factors computed in our study was motivated by previous research reported in Hinterleitner et al. (2011a, b, 2012), as these were very closely related to the objectives of our study. The interpretation of the factors involves exploring the indicator-factor relationships by investigating the factor loadings. In Hair et al. (2009), authors define a practically significant cut-off threshold of 0.5 for a factor loading to be significant and the indicators that loads at 0.5 or higher on two or more factors are considered cross-loaders. Therefore, in this study a threshold of 0.5 was used for factor loadings to interpret indicator-factor relationships. Moreover, towards establishing a more reliable factor structure, EFA was also performed on random subsamples of data extending from $N = 165$ to $N = 264$ with increments of 2. This exploratory analysis allowed us to vary the sample to variable ratios from 15:1 to 24:1, thus further validating the data sufficiency hypothesis. The key steps involved in performing EFA are summarised in Fig. 1.

Confirmatory factor analysis

The EFA forms the conceptual and theoretical foundation for the factor models describing 'indicator-latent factor' relationships. The confirmatory factor analysis (CFA), in turn, explicitly and directly tests the 'fit' of the factor model developed using EFA (Thompson 2004). The CFA requires researchers to have specific expectations regarding the number of factors, indicator-latent factor relationships, and the correlation between the latent factors, thus an established theory is needed for CFA. The CFA allows for

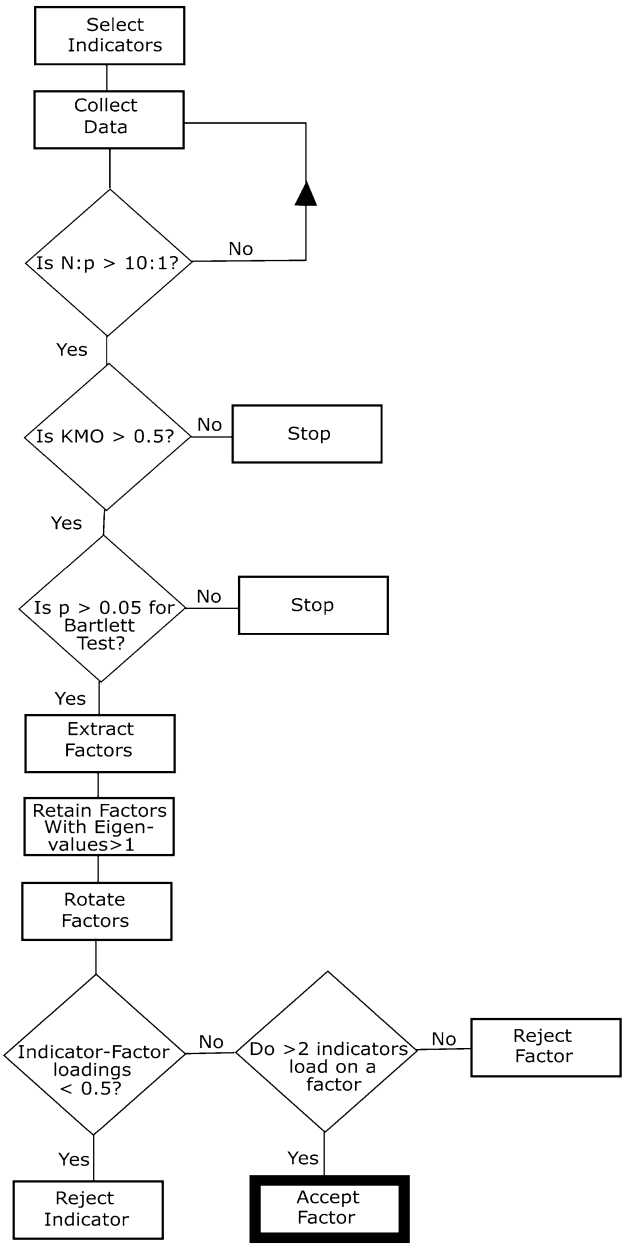

Fig. 1 The key steps involved in exploratory factor analysis

the direct testing of theory and quantifying the degree of model fit.

Formulation

The CFA model can be expressed as follows (Anderson and Gerbing 1988; Vandenberg and Lance 2000):

$$x = \tau + \Lambda\xi + \delta, \quad (1)$$

where x is a vector of 'n' indicators, τ is a vector of 'n' intercepts, ξ is a vector of 'i' latent factors such that $i < n$, Λ is a $n \times i$ matrix of factor loadings that relate indicators to the latent factors, and δ is a vector of 'n' variables that represent random errors in measurement and measurement specificity of the indicators. In most CFA applications, the intercepts are assumed to be zero and are not estimated (Vandenberg and Lance 2000). The model also assumes that the $E(\xi\delta) = 0$ and that the variance-covariance matrix for x, denoted as Σ, is given by:

$$\Sigma = \Lambda\Phi\Lambda' + \Theta, \quad (2)$$

where Φ is the $i \times i$ matrix of ξ and Θ is the diagonal $n \times n$ covariance matrix of δ.

Methodology

CFA mainly concerns with modelling the latent factors that account for commonality among the set of indicators. The commonality between measures of a construct can be depicted using path diagrams (Hoyle 2000). In path diagrams, the measured variables or indicators are represented using rectangles and the unmeasured variables by ellipses. As such, latent factors are represented using large ellipses and unobserved measurement errors that affect indicators as smaller ellipses. The causality relationships are indicated using a single headed arrow, whereas double-headed curved arrows are used to represent variances. Two different models exist—principal factor (reflective) and composite latent variable (formative)—to describe the causality relationships between latent factors, indicators and errors of measurement (Jarvis et al. 2003). The reflective model expects the latent factors to cause changes in the indicators, whereas in the formative model the indicators are expected to affect changes in the latent factors (Jarvis et al. 2003). The decision rules or guiding principles to choose the appropriate model are listed in Jarvis et al. (2003). Based on such rules, the reflective model is shown to be better suited for the current study. Therefore, the indicators were expected to be caused by two unmeasured influences: (1) a causal relationship they share with other indicators (i.e., the latent factor), and (2) a causal influence unique to each indicator that is quantified using the errors of measurement (Hoyle 2000).

There are a variety of statistical packages available for implementing CFA, such as MPlus (Byrne 2013a), AMOS (Byrne 2013b), and lavaan (Rosseel 2012). For the current study, we have implemented CFA using the lavaan (Latent Variable Analysis) package for R. The lavaan package allows the specification of the CFA model (as implemented in the path diagram) through the model syntax. The model syntax is a description of the model that needs to be estimated. The lavaan package allows estimates of various goodness-of-fit measures for the developed model, as detailed next.

Goodness-of-fit metrics

The factor model is considered acceptable if the covariance structure implied by the model matches the covariance structure of the sampled data (Cheung and Rensvold 2002). The acceptability of the model is reflected in its goodness-of-fit (GOF) index. The most common GOF index is the 'χ^2' metric that measures the GOF derived from the fitting function that measures the relationship between the observed and the implied covariance matrices. The 'χ^2' metric tests the null hypothesis of 'χ^2' being equal to 0, which indicates the best possible fit (Cheung and Rensvold 2002). The 'χ^2' test, however, is greatly affected by sample size (Cheung and Rensvold 2002). Therefore, other GOF indices have been proposed previously, such as the comparative fit index (CFI), normed fit index (NFI), non-normed fit index (NNFI), incremental fit index (IFI), relative non-centrality index (RNI), goodness-of-fit index (GFI), and standardized root mean square residual (SRMR) (Jackson et al. 2009; Cheung and Rensvold 2002). The CFI, NFI, NNFI, IFI and RNI indices compare the performance of the model with a baseline (or null) model that assumes zero correlation between all the indicators. The GFI, on the other hand, does not compare the model to a baseline model and is computed based on the amount of variance explained by the model. Finally, the SRMR index is estimated by computing the mean absolute value of the covariance of residuals. Typically, values $\geq$0.90 are considered adequate for the CFI, NFI, NNFI, IFI, RNI and GFI indices (Bagozzi and Yi 1988; Bentler and Bonett 1980), whereas a value of SRMR $\leq$0.08 (Vandenberg and Lance 2000) reflects the adequate fit of a model. Here, a combination of these indices is used for model validation.

Measurement and structural invariance

The CFA forms a part of larger family of structural equation modelling (SEM) methods. The SEM methods are a

broad class of statistical models that consist of two parts: the measurement model and the structural model (Jackson et al. 2009; Beaujean 2014). The measurement model reflects the relationship between the latent factors and the indicators, whereas the structural model relates the relationship of latent factors to each other (Jarvis et al. 2003). Towards establishing the reliability and validity of the measurement and structural model, it is important to establish the between-group invariance (or equivalence) of the models (Vandenberg and Lance 2000). The measurement and structural invariance of the model help verify: (1) the conceptual equivalence of the latent factors across groups, and (2) the equivalence of associations between indicators and factors and between factors across groups. The invariance of models is demonstrated by testing a number of hypotheses regarding measurement and structural invariance (Vandenberg and Lance 2000).

The first hypothesis tests for the equivalence of the pattern of zero and non-zero coefficients in the matrix of factor loadings (Λ in Eq. 1) (Oort 2005). The hypothesis is tested by estimating the same model for each group simultaneously while allowing estimated parameters to differ. The hypothesis tests for the equivalence of the models through a χ^2 test. Therefore, a p value ≤ 0.05 rejects the hypothesis of both the models being equivalent; however, a p value greater than 0.05 leads to configural invariance (Beaujean 2014).

The second hypothesis tests for the equivalence of the unstandardized factor loadings across groups Sass (2011) by constraining loadings to be equal between groups and is referred to as metric or weak invariance. An additional test evaluates the equivalence of unstandardized intercepts or thresholds across groups by constraining intercepts to be equal between groups, and is called scalar or strong invariance. An alternate test evaluates the equivalence of residuals across groups by constraining error variances to be equal between groups and is known as uniqueness or strict invariance (Beaujean 2014). Combined, the configural, metric, scalar and strict invariances evaluate the measurement invariance of the model as these steps are mainly concerned with the indicator-latent variable relationships (Beaujean 2014). Structural invariance testing, on the other hand, evaluates the properties of latent variables, thus involves constraining variances, covariances and means of the latent factors in a stepwise manner.

If the level of invariance for all variables is untenable, a follow-up analysis is needed to determine which indicators are contributing to model misfit. This follow-up analysis involves invariance testing while leaving the non-invariant indicators in the model and not constraining them to be invariant across the groups. The resulting invariance model is said to have partial invariance, that warrants invariance for most of the parameter estimates with the exception of a few parameters within an invariance model (Beaujean 2014). The non-invariant indicators are identified using their modification indices. The modification index estimates the amount of overall decrease in the χ^2 value if the previously constrained parameter was freely estimated (Kline 2013). The modification index is interpreted as the χ^2 statistic with a single degree of freedom (Kline 2013).

Measurement and structural invariance of the model can be interpreted using the response shift theory (Oort 2005; Sass 2011; de Beurs et al. 2015). The response shift is defined as: "a change in the meaning of one's self-evaluation of a target construct as a result of (a) a change in the respondent's internal standards of measurement (i.e., scale recalibration); (b) a change in the respondent's values (i.e., the importance of component domains constituting the target construct through reprioritization) or (c) a redefinition of the target construct (i.e., reconceptualization)" (Schwartz and Sprangers 1999). The concepts represented by the factors are reflected in the patterns of zero and non-zero factor loadings in the Λ matrix (Oort 2005). Therefore, according to response shift theory a configural non-invariance that leads unequal factor loading patterns across groups, occurs due to reconceptualization. The reconceptualization reflects a change in the meaning of the indicators and, thus, leading to change in the conceptual representation of the latent factors (Barclay and Tate 2014). Furthermore, the metric non-invariance occurs due to reprioritization that involves an indicator becoming more or less indicative of a concept (Oort 2005). The graphical representation of the reprioritization is shown in Fig. 2 indicating underestimation of the indicator values for a group with lower loading values, regardless of the value of latent construct/factor (Wicherts and Dolan 2010). For example, let us assume that one of the latent factors for the present study is listening pleasure and the indicator that shows reprioritization across natural and synthesised voices is acceptance with $\lambda_{nat} > \lambda_{tts}$, in this case it can be said that the acceptance natural voices will be higher as compared to synthesised voices irrespective of the listening pleasure they offer. The scalar and strict non-invariance, in turn, represent uniform and non-uniform recalibration (Oort 2005). The recalibration process indicates a change in the internal standards of the participants and if the change affects all response options in the same direction and to the same extent then it leads to uniform recalibration (Oort 2005). The graphical representation of the uniform recalibration is shown in Fig. 3 indicating underestimation of the indicator values for a group with lower intercept values, regardless of the value of latent construct/factor (Wicherts and Dolan 2010). Moreover, a non-invariant factor variance model suggests true changes in the variances of the

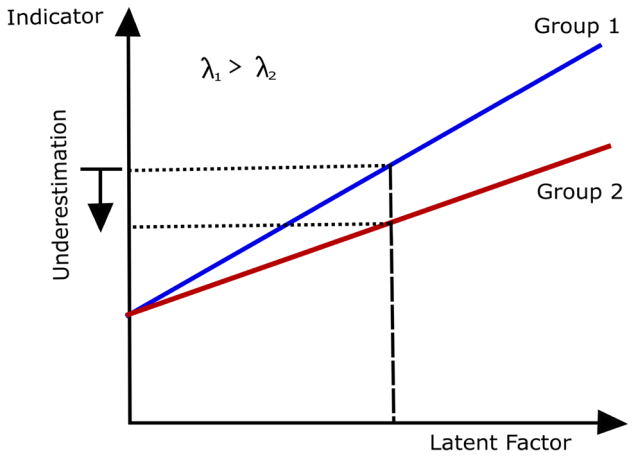

Fig. 2 Graphical representation of reprioritization

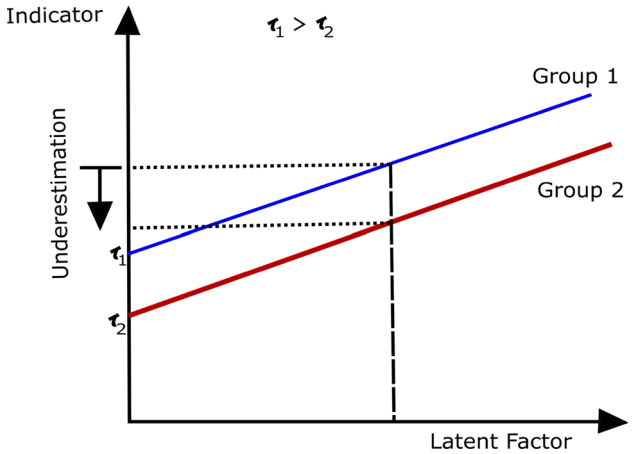

Fig. 3 Graphical representation of recalibration

factor, whereas a non-invariant factor covariance model indicates higher level reconceptualization or reprioritization (Oort 2005). Finally, a non-invariant factor means the model reflects true changes in the factor means across groups (Oort 2005). The key steps involved in performing CFA followed by measurement and structural invariance tests are summarised in Fig. 4.

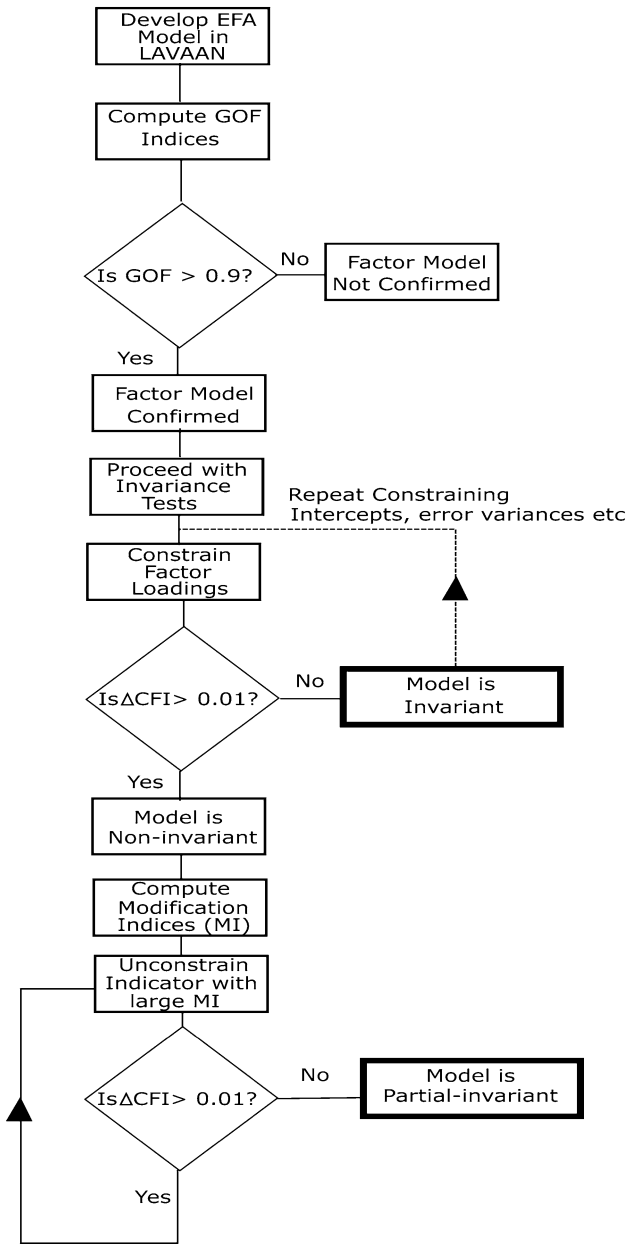

Fig. 4 The key steps involved in confirmatory factor analysis and invariance tests

Results

Exploratory factor analysis

The data was first tested for internal consistency reliability using Cronbach's alpha (α) and a value of $\alpha = 0.89 \pm 0.03$ was obtained, thus establishing the reliability of the ratings used in the study. Following that, Kaiser–Meyer–Olkin (KMO) measure of sampling adequacy along with Bartlett's test of sphericity were computed to assess the adequacy correlation matrix for factor analysis. The measured KMO for the data was 0.94 thus supporting factor analysis. The Bartlett's test of sphericity resulted in significance levels below 0.05, thus confirming significant relationships between ratings. These measures established the adequacy of the data for exploratory factor analysis. As a next step, the number of factors necessary for EFA was obtained. Both Kaiser's criterion and the Scree test recommended that two factors be retained (factor 1: eigenvalue = 7.251; factor 2: eigenvalue = 1.05).

Next, EFA was performed using principal axis factoring along with promax rotation to reduce cross loadings of

items. This resulted in two factors, where the ratings *voice pleasantness*, *acceptance*, *listening effort*, *comprehension problems* and *valence* loaded on most significantly to Factor 1. Ratings *intonation*, *emotion*, and *arousal*, in turn, loaded on to Factor 2, as shown in Table 3. To obtain meaningful factors, factor loadings below 0.5 were not considered. Therefore, dominance and speaking rate did not load significantly on any of the factors. Furthermore, naturalness dimension was not utilised for further analysis as it showed high cross-loadings between the two latent constructs. Also, as visible from Table 3, the sub-sampling factor analysis, employed to validate the reliability of the factor structure and data sufficiency, produced similar factor structure and mean loadings along with very low standard deviations over the obtained factor loadings. Moreover, the cumulative variance explained by the two factors was 57%.

Confirmatory factor analysis

Towards verifying the factor structure obtained from EFA, a confirmatory factor analysis or CFA was performed. The model fit parameters obtained from CFA, as reported in Table 4, validate the model as the fit parameters GFI, NFI, NNFI, CFI, RNI and IFI were observed to be greater than 0.90 and SRMR was found to be less than 0.08. Following the CFA, measurement invariance (MI) and structural invariance (SI) for the model were examined for different groups in the data. First, invariance tests were performed between samples from groups of *female and male* participants, followed by samples from groups of *natural and synthesized* speech stimuli.

Table 3 Factor loadings obtained for each item using EFA

Rating	General EFA		Subsampling EFA			
	Factor Loadings		Mean		Std. Dev.	
	1	2	1	2	1	2
VP	**0.85**	0.14	**0.847**	0.140	**0.008**	0.009
Acc	**0.80**	0.15	**0.798**	0.153	**0.007**	0.008
LE	**0.91**	0.03	**0.899**	0.043	**0.010**	0.012
CP	**−0.73**	0.13	**−0.723**	0.116	**0.012**	0.013
Val	**0.84**	0.09	**0.840**	0.092	**0.008**	0.009
Int	0.17	**0.74**	0.170	**0.742**	0.012	**0.012**
Emo	−0.12	**1.02**	−0.109	**1.013**	0.011	**0.012**
Ar	0.27	**0.52**	0.274	**0.513**	0.012	**0.011**
Nat	0.45	0.52	0.449	0.520	0.006	0.006
SR	−0.20	−0.10	−0.198	−0.102	0.012	0.013
Dom	0.34	0.33	0.335	0.335	0.012	0.013

Bold indicates values greater than 0.5 are significant (>0.5)

Table 4 Goodness-of-fit metrics obtained using CFA

χ^2	df	GFI	NFI	NNFI	SRMR	CFI	RNI	IFI
152.10	19	0.984	0.980	0.979	0.032	0.981	0.981	0.981

MI and SI testing for female and male groups

The results from MI and SI tests between male and female listeners are reported in Table 5. The configural invariance model (model 1) proved to have a good fit judging by its CFI value of 0.979 and SRMR of 0.03. The metric invariance model (model 2) also resulted in CFI value of 0.980 and SRMR value of 0.033. Moreover, the $\Delta\chi^2$ value between configural and metric invariance model was 3.26 with *p value* of 0.836 which reflects insignificant difference between the two models. Thus, it is evident that the developed model was metric invariant between female and male raters. Similarly, the $\Delta\chi^2$ value between metric invariance and scalar invariance (model 3) models was 9.85 with *p value* of 0.131, indicating scalar invariance of the model. Finally, comparing the scalar invariance model and the strict invariance model (model 4) the $\Delta\chi^2$ value was found to be equal to 8.66 with *p value* of 0.372, thus indicating no significant difference between the models and establishing the strict invariance of the proposed model.

Moreover, the SI tests involved developing models with constrained latent variables' variances (model 5), covariances (model 6) and means (model 7) sequentially. The developed models were compared against the preceding model to test for structural invariance. It is important to note that comparisons between models 4 and 5 and between models 5 and 6 resulted in χ^2 tests showing non-invariance; however, as the ΔCFI values were lower than 0.01, models 5 and 6 were invariant. From the insignificant differences between the model, as reported in Table 5, it is evident that the developed models proved existence of structural invariance of the model between male and female listeners.

However, it should be noted that the there was imbalance in the number of female and male voices and listeners. Therefore, we implemented the measurement and structural invariance tests with gender balanced listeners (8 female and 8 randomly selected males) and synthesised voice only. The results further indicated measurement and structural invariance between male and female listeners.

MI testing for TTS and natural speech stimuli groups

The results from MI and SI tests between natural and synthesized voices are reported in Table 6. The configural

Table 5 Measurement and structural invariance testing for groups of female and male raters

Invariance	Model	χ^2	df	CFI	GFI	NFI	NNFI	SRMR	RNI	IFI	Model comparison	$\Delta\chi^2$	Δ df	p value	ΔCFI
Measurement	1	184.80	38	0.979	0.987	0.974	0.970	0.030	0.979	0.979	–	–	–	–	–
	2	187.57	44	0.980	0.987	0.974	0.974	0.033	0.980	0.980	1 vs. 2	2.77	6	0.836	0.001
	3	197.43	50	0.979	0.987	0.973	0.977	0.033	0.979	0.979	2 vs. 3	9.85	6	0.131	0.001
	4	206.09	58	0.979	0.986	0.971	0.980	0.033	0.979	0.979	3 vs. 4	8.66	8	0.372	0
Structural	5	199.41	59	0.980	0.987	0.972	0.981	0.039	0.980	0.980	4 vs. 5	6.68	1	<0.01	0.001
	6	206.12	60	0.980	0.986	0.971	0.981	0.037	0.980	0.980	5 vs. 6	6.71	1	<0.01	0
	7	208.95	62	0.979	0.986	0.971	0.981	0.039	0.979	0.979	6 vs. 7	2.82	2	0.243	0.001

invariance model (model 1) resulted in a well fitted model as its CFI value and SRMR values were 0.961 and 0.048, respectively. The metric invariance model (model 2) was then developed by constraining the factor loadings across the two groups. The configural invariance and metric invariance models were then compared using the χ^2 test, which resulted in $\Delta\chi^2$ value of 65.53 with p value less than 0.01 and ΔCFI greater than 0.01. Thus, suggesting metric non-invariance for the developed model across natural and TTS generated speech stimuli. The source of non-invariance was extracted by exploring the modification indices. The subjective dimensions of 'Comprehension Problems' (CP) and 'Arousal' (Ar) resulted in significantly high modification indices. Thus, allowing CP and Ar to vary in an unconstrained manner, a partial metric invariance model (model 2a) was developed. The comparison between partial metric invariance model and configural invariance model resulted in $\Delta\chi^2$ value of 30.54 with p value less than 0.037, suggesting partial metric non-invariance. However, the ΔCFI value between the two models was 0.006, which is less than 0.01, thus indicating partial metric invariance according to the recommendations in Cheung and Rensvold (2002), which suggests using $\leq \Delta CFI = 0.01$ value as a better reflection of model invariance compared to χ^2 test. The unconstrained loading values for CP and Ar indicators, for partially metric invariant model, are reported in Table 7.

Next, a scalar invariance model (model 3) was developed by constraining the model intercepts and it was compared with the partial metric invariant model developed above. The resulting $\Delta\chi^2$ value was 41.72 with p value less than 0.01 and ΔCFI greater than 0.01, which suggests a significant difference between the two models resulting in scalar non-invariance. The exploration of modification indices suggested 'Acceptance' as the source of scalar non-invariance. Thus, in the following model Acc intercepts were unconstrained and a partial scalar invariant model (model 3a) was developed. For validating the model, a χ^2 test between partial metric invariant and partial scalar invariant models resulted in ΔCFI value less than 0.01 thus, suggesting partial scalar invariance of the model. The unconstrained intercept values for Acc indicators, for partially scalar invariant model, are reported in Table 7.

Moreover, a strict invariance model (model 4) was implemented by constraining the model residuals across the two groups. The comparison between the partial scalar

Table 6 Measurement and structural invariance testing for groups of natural and synthesized speech samples

Invariance	Model	χ^2	df	CFI	GFI	NFI	NNFI	SRMR	RNI	IFI	Model Comparison	$\Delta\chi^2$	Δ df	p value	ΔCFI
Measurement	1	218.56	38	0.961	0.987	0.953	0.942	0.048	0.961	0.961	–	–	–	–	–
	2	284.09	44	0.948	0.983	0.939	0.934	0.076	0.948	0.948	1 vs. 2	650.53	6	<0.01	0.013
	2a	249.10	42	0.955	0.985	0.947	0.940	0.058	0.955	0.955	1 vs. 2a	300.54	4	<0.01	0.006
	3	290.82	48	0.944	0.983	0.938	0.939	0.068	0.944	0.944	2a vs. 3	410.72	6	<0.01	0.011
	3a	270.58	47	0.952	0.984	0.942	0.942	0.065	0.952	0.952	2a vs. 3a	210.48	5	<0.01	0.003
	4	397.58	55	0.926	0.978	0.915	0.924	0.085	0.926	0.926	3a vs. 4	127	8	<0.01	0.026
	4a	318.33	54	0.943	0.981	0.932	0.941	0.068	0.943	0.943	3a vs. 4a	470.75	7	<0.01	0.009
Structural	5	318.95	56	0.943	0.981	0.932	0.943	0.071	0.943	0.943	4a vs. 5	0.62	2	0.73	0
	6	319.57	57	0.943	0.981	0.932	0.944	0.070	0.943	0.943	5 vs. 6	0.62	1	0.43	0
	7	941.14	59	0.809	0.942	0.799	0.819	0.597	0.809	0.809	6 vs. 7	6210.57	2	<0.01	0.134

Table 7 Unconstrained parameter values for partial metric and scalar invariant models.

Invariance Model	Parameter	Indicators	Natural	Synthesized
Metric	Loadings (λ)	CP	−0.36	−0.76
		Ar	0.55	0.39
Scalar	Intercepts (τ)	Acc	0.72	0.65

invariant model and strict invariant model resulted in $\Delta\chi^2$ value was 127 with p value less than 0.01 thus, suggesting strict non-invariance of the model. The source of non-invariance was 'Comprehension Problems' due to high modification index. The partial strict invariant model (model 4a) was then implemented by allowing CP residuals to vary freely across groups. The comparison between partial strong invariant and partial strict invariant models resulted in $\Delta\chi^2$ value of 47.75 with p value less than 0.01 thus, suggesting partial non-invariance. However, the ΔCFI value between the two models was 0.009, which is less than 0.01 thus, suggesting partial strict invariance according to the recommendations in (Cheung and Rensvold 2002).

The SI tests were then performed by developing models with constrained latent factor variances, covariances and means, and comparing them with preceding models. The models with constrained latent factor variances (model 5) and covariance (model 6) showed invariance as the differences between the developed models and the preceding models were insignificant. However, the model with constrained factor means (model 7) was evidently significantly different compared to previous model that reflects non-invariance.

Discussion

Exploratory factor analysis

EFA was performed using all the subjective dimensions except overall impression, as it comprises information from other dimensions (Hinterleitner et al. 2011b). The EFA resulted in extraction of two factors, with *factor 1* with loadings from voice pleasantness, acceptance, listening effort, comprehension problems and valence, and *factor 2* with loadings from intonation, emotion, and arousal. Thus, it is evident that the items which load on *factor 1* cover the *listening pleasure* and intelligibility of the systems, whereas items which load on to *factor 2* reflect the signal *prosody* and rhythm. Moreover, the sub-sampling EFA validated the obtained factor structure as it resulted in similar factor loadings with low variations (given by standard deviation) for each indicator. This indicates towards the existence of two perceptual dimensions namely, 'listening pleasure' and 'prosody' for the measurement of QoE for synthesised speech developed for personal digital assistants. Furthermore, the findings are in corroboration with exploratory factor analysis performed for audiobooks, as reported in Hinterleitner et al. (2011b), with the exception that in the current study the 'comprehension problems' scale negatively loaded on factor 1, rather than cross-loading on both the factors as reported in Hinterleitner et al. (2011b). This suggests inverse relationship between comprehension problems and 'listening pleasure and intelligibility' for personal digital assistant systems.

Previous research using EFA and multidimensional scaling for extracting the perceptual dimensions of synthetic speech QoE has found three major dimensions namely, naturalness of voice, prosodic quality and, fluency and intelligibility (Norrenbrock 2015). However, for majority of the tests reported in Norrenbrock (2015), naturalness cross-loaded on the first two dimensions, which is consistent with the findings from our study. This renders the task of attributing naturalness to a particular perceptual dimension difficult. Therefore, for model simplification the naturalness scale was not considered in further analysis.

Furthermore, the valence and arousal scales loaded on two different factors, factor 1 and 2, respectively. The valence and arousal scales form the two orthogonal dimensions of the emotional/affective experience corresponding to positiveness/pleasantness and alertness (Tseng 2014), respectively. Thus, the loading of valence item on factor 1 further establishes relationship of factor 1 to perceptual dimension of 'listening pleasure'. Also, the loading of arousal scale on factor 2 relates stimulus evoked alertness to prosody in speech, which is also corroborated by previous findings reported in Syrdal and Kim (2008). These findings indicate that changes in underlying perceptual constructs of QoE, due to changes in system quality, alters users' affective states. Therefore, it is evident that affective scales corresponding to valence and arousal dimensions are important for estimating underlying perceptual dimensions of users' experience with personal digital assistants.

However, another model for users' affect utilizes an additional dimension of the so-called dominance for describing users' control over a situation (Bradley and Lang 1994). The dominance scale did not load significantly on to any of the two extracted factors. Similarly, the speaking rate scale did not show any significant loadings

on either of the two factors. The insignificant factor loadings for dominance and speaking rate can be attributed to their low F-statistic values obtained using ANOVA, as reported in Gupta et al. (2015), thus suggesting low inter-class variation compared to intra-class variation for these scales. Therefore, both dominance and speaking rate scales were rejected from further analysis.

The EFA established the factor model for confirmatory factor analysis, as shown in the path diagram depicted by Fig. 5. The path diagram suggests indicates that the presented model is reflective, i.e., the latent factors (such as, listening pleasures and prosody) cause changes in the indicators (such as, voice pleasantness and intonation). The factor model consists of two correlated latent factors as these were estimated using a promax rotation that results in oblique (non-orthogonal) factors. The first and second factors can be measured using five and four continuous factor indicators, respectively. The loadings from factor indicators and previous research (Hinterleitner et al. 2011b) suggest the first and second factor represent listening pleasure and prosody, respectively.

Confirmatory factor analysis

The factor model for evaluating the perceptual dimensions of the synthesized speech QoE was confirmed using the confirmatory factor analysis as all the model fit parameters satisfied the goodness-of-fit criteria. This suggests that the used items serve as good indicators of the underlying perceptual dimensions. However, measurement invariance of the developed model needs to be established for it be useful while applying it to various groups, such as male and female listeners or natural and synthesized speech. Therefore, the measurement and structural invariance of the model were tested following CFA.

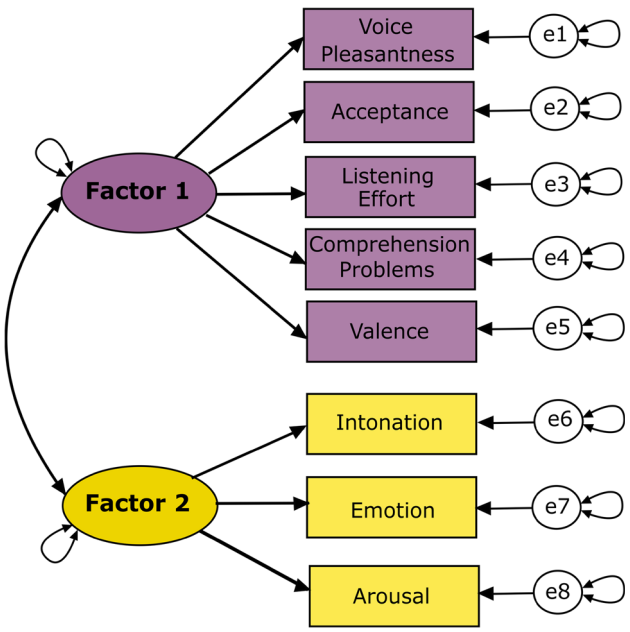

Fig. 5 The factor model for confirmatory factor analysis

One of the limitations of this study is that it leverages data formed from multiple responses from each individual listener, while each response being treated as an independent observation, which has been criticized in Viswanathan and Viswanathan (2005). However, given the time and monetary costs of conducting auditory listening tests it is difficult to obtain data from more subjects. Moreover, there have been previous EFA studies (Hinterleitner et al. 2011a, b), along with the EFA studies reported in Viswanathan and Viswanathan (2005), which treated multiple responses from individual listeners as independent observations, and concluded with similar factor structures. Furthermore, it is important for the listening tests to include a wide spectrum of TTS systems to extract more generalizable results (Hinterleitner et al. 2012). Hence, in this study we utilized speech stimuli from four different natural voices and seven different PDA systems, based on different TTS systems.

Measurement and structural invariance for male vs. female listeners

The model was found to be measurement and structurally invariant between male and female listeners. The MI established that the indicators measured similar latent factors/constructs across groups whereas, the SI tests validated the reliability of the obtained latent factors/constructs. This indicates that the developed model can be used consistently across raters to gauge listening pleasure and prosodic information of natural or synthetic speech. However, it should be noted that this study did not incorporate any natural male voices and therefore, further work should evaluate the effects of natural male voices talents on latent constructs.

Measurement invariance for natural vs. synthesized speech stimuli

The model was also tested for measurement invariance between the groups of natural and synthesized speech stimuli. The model showed weak non-invariance, which can be attributed to the 'comprehension problems' and 'acceptance' items. Allowing the CP and Ar items to vary freely led to partial weak invariance in the model. For the partially weak invariant model the factor loading of CP was higher for synthesized speech in comparison to natural speech, thus indicating reprioritization of CP for synthesized speech. The reprioritization of CP suggests that CP is more indicative of listening pleasure of synthesized speech

(Oort 2005). This is further elaborated in Fig. 6 and Table 7, where the change in listening pleasure for a stimulus leads to higher change in CP for synthesized speech compared to natural speech stimulus. Furthermore, the factor loading of Ar was higher for natural speech in comparison to synthesized speech, thus indicating reprioritization of Ar for natural speech. This indicates that Ar scale is more relevant for natural speech than synthesized speech while evaluating prosody Oort (2005). It is further evident from Fig. 7 and Table 7, where a change in prosody leads to higher change in Ar of natural speech compared to synthesized speech.

Next, the model was tested for strong invariance that suggested partial strong invariance. The origin of strong non-invariance was found to be the 'acceptance' items, thus suggesting recalibration (Oort 2005) of this item between natural and synthesized speech stimuli. The intercept for natural speech was higher compared to synthesized speech for 'acceptance' scale as evident from Fig. 8 and Table 7. This indicates that the acceptance of natural speech is higher than synthesized speech regardless of change in 'prosody', thus suggesting a listener bias while scoring the speech naturalness. Moreover, this indicates that to achieve the acceptance level of natural speech changing the listening pleasure of synthesized speech is insufficient.

Furthermore, a partial strict invariance was observed for the model. The source of strict non-invariance was found to be 'comprehension problems', thus suggesting non-uniform recalibration (Oort 2005) of 'comprehension problems' between natural and synthesized speech. The non-uniform recalibration indicates that the response shifts for individual subjects were in different directions leading to changes that appear in the variances of the residual factors (Oort 2005). The measurement invariance tests indicated

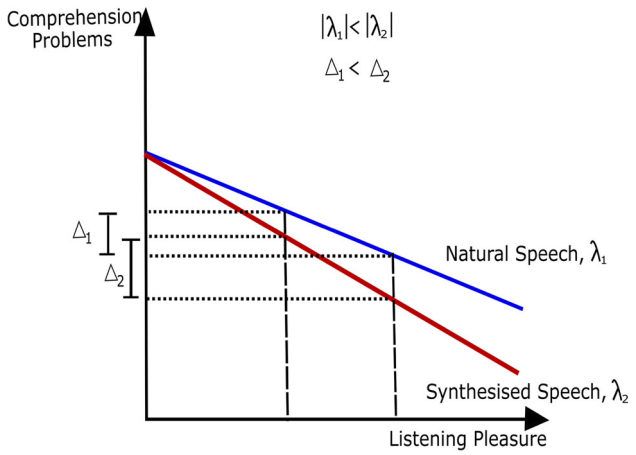

Fig. 6 Graphical representation of reprioritization in comprehension problems indicator across natural and synthesized speech, based on Table 7

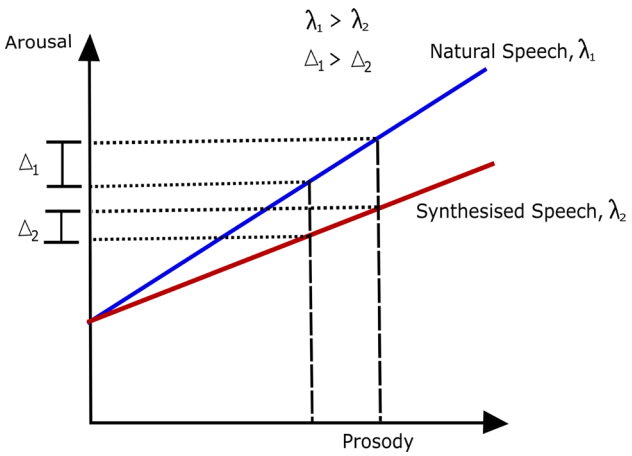

Fig. 7 Graphical representation of reprioritization in arousal indicator across natural and synthesized speech, based on Table 7

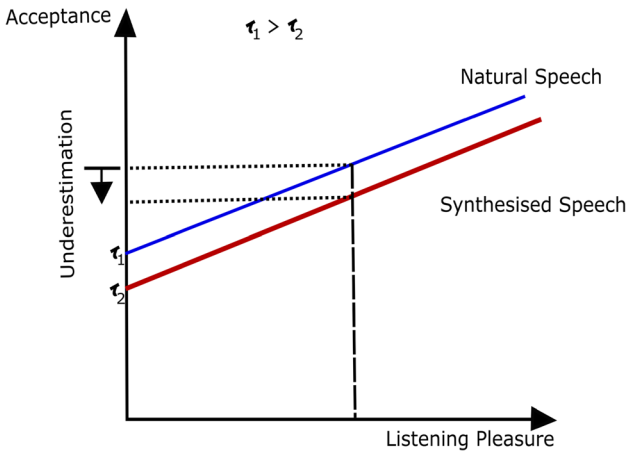

Fig. 8 Graphical representation of recalibration in acceptance indicator across natural and synthesized speech, based on Table 7

that the indicators used in the experiment measured similar latent factors between synthesized and natural speech samples, thus validating the reliability of the indicators.

Finally, the structural invariance tests indicated that the models with constrained latent factor variances and covariances were invariant. However, the model with constrained factor means showed non-invariance thus, reflecting true changes in the factor means. The true change reflects change in listeners' level of the target latent factors between the two groups (Oort 2005). Therefore, evidently natural and synthetic voices would score differently on listening pleasure and prosody constructs.

Conclusion

The present study involved an auditory listening test where the listeners were asked to score their perceived QoE of natural and synthesized voices on various indicators, such

as voice pleasantness, listening effort and appropriateness of intonation. Following that, exploratory and confirmatory factor analyses were conducted to extract perceptual dimensions or latent factors of the quality space. The two extracted factors were 'listening pleasure' and 'prosody,' thus corroborating previous findings based on audiobooks. Next, a model was developed incorporating relationships between the indicators and the perceptual dimensions. The developed model was tested for invariance or equivalence across groups of male and female listeners, as well as natural and synthetic voices. The invariance tests established the conceptual equivalence of the obtained perceptual dimensions across the different groups. Therefore, in future studies involving the evaluation of natural voice talents for the development of TTS systems for PDAs or synthesised voices for PDAs, the nine indicators listed in the study can be used to measure their listening pleasure and prosody information. However, a model to measure the QoE from the latent factors needs to be developed in the future work.

Acknowledgements The authors thank MDEIE, FQRNT, and NSERC for funding; H. J. Banville, R. Cassani, A. Clerico, and I. Albuquerque for help with data acquisition; and Nuance Communications for invaluable discussions and access to relevant voice talent recordings. Funding was provided by NSERC (Grant no RGPIN 402237-2011) and FQRNT (Grant no 2014 -NC-173415).

References

Anderson JC, Gerbing DW (1988) Structural equation modeling in practice: a review and recommended two-step approach. Psychol Bull 103(3):411

Bagozzi RP, Yi Y (1988) On the evaluation of structural equation models. J Acad Mark Sci 16(1):74–94

Barclay R, Tate RB (2014) Response shift recalibration and reprioritization in health-related quality of life was identified prospectively in older men with and without stroke. J Clin Epidemiol 67(5):500–507

Bartlett MS (1950) Tests of significance in factor analysis. Br J Stat Psychol 3(2):77–85

Beaujean AA (2014) Latent variable modeling using R: a step-by-step guide, Routledge

Bentler PM, Bonett DG (1980) Significance tests and goodness of fit in the analysis of covariance structures. Psychol Bull 88(3):588

Black AW, Taylor P (1994) CHATR: a generic speech synthesis system. In: Proceedings of the 15th conference on Computational linguistics, vol 2. Association for Computational Linguistics, Stroudsburg, PA, USA, pp 983–986

Bradley M, Lang P (1994) Measuring emotion: the self-assessment manikin and the semantic differential. J Behav Ther Exp Psychiatry 25(1):49–59

Brunnström K, Beker SA, De Moor K, Dooms A, Egger S, Garcia MN, Hossfeld T, Jumisko-Pyykkö S, Keimel C, Larabi MC, Lawlor B (2013) Qualinet White Paper on Definitions of Quality of Experience Output from the fifth Qualinet meeting, Novi Sad, Version 1.2, Technical report, Qualinet COSTIC 1003

Byrne BM (2013a) Structural equation modeling with Mplus: basic concepts, applications, and programming, Routledge

Byrne BM (2013b) Structural equation modeling with AMOS: basic concepts, applications, and programming, Routledge

Cheung GW, Rensvold RB (2002) Evaluating goodness-of-fit indexes for testing measurement invariance. Struct Equ Model 9(2):233–255

Comrey AL, Lee HB (2013) A first course in factor analysis. Psychology Press, 2nd edn. Erlbaum, Hillsdale, NJ

Costello AB, Osborne JW (2005) Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. Pract Assess Res Eval 10:173–178

de Beurs DP, Fokkema M, de Groot MH, de Keijser J, Kerkhof AJ (2015) Longitudinal measurement invariance of the Beck Scale for Suicide Ideation. Psychiatry Res 225(3):368–373

Fabrigar LR, Wegener DT, MacCallum RC, Strahan EJ (1999) Evaluating the use of exploratory factor analysis in psychological research. Psychol Methods 4(3):272

Gupta R, Banville HJ, Falk TH (2015) PhySyQX: a database for physiological evaluation of synthesised speech quality-of-experience. In: IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA). IEEE, pp 1–5. doi:10.1109/WASPAA.2015.7336888

Hair JF, Black WC, Babin BJ, Anderson RE (2009) Multivariate data analysis, vol 7. Pearson Prentice Hall, Upper Saddle River

Henson RK, Roberts JK (2006) Use of exploratory factor analysis in published research common errors and some comment on improved practice. Educ Psychol Meas 66(3):393–416

Hinterleitner F, Möller S, Norrenbrock C, Heute U (2011a) Perceptual quality dimensions of text-to-speech systems. In: Proceedings of the Twelfth Annual Conference of the International Speech Communication Association (Interspeech), Florence, Italy, pp 2177–2180

Hinterleitner F, Neitzel G, Möller S, Norrenbrock C (2011b) An evaluation protocol for the subjective assessment of text-to-speech in audiobook reading tasks. In: Proceedings of the Blizzard challenge workshop, International Speech Communication Association (ISCA), Florence, Italy

Hinterleitner F, Norrenbrock C, Moller S, Heute U (2012) What makes this voice sound so bad? A multidimensional analysis of state-of-the-art text-to-speech systems. In: Spoken Language Technology Workshop (SLT). IEEE, pp 240–245. doi:10.1109/SLT.2012.6424229

Hinterleitner F, Norrenbrock C, Moller S, Heute U (2014) Text-to-speech synthesis. In: Quality of experience, pp 179–193. doi:10.1007/978-3-319-02681-7_13

Hoyle RH (2000) Confirmatory factor analysis. In: Tinsely HEA, Brown SD (eds) Handbook of applied multivariate statistics and mathematical modeling. Academic press, New york, pp 465–497

ITU-T, P. 85 (2016) A Method for Subjective Performance Assessment of the Quality of Speech Voice Output Devices, International Telecommunication Union, CH-Genf

Jackson DL, Gillaspy JA Jr, Purc-Stephenson R (2009) Reporting practices in confirmatory factor analysis: an overview and some recommendations. Psychol Methods 14(1):6

Jarvis CB, MacKenzie SB, Podsakoff PM (2003) A critical review of construct indicators and measurement model misspecification in marketing and consumer research. J Consum Res 30(2):199–218

Kaiser HF (1960) The application of electronic computers to factor analysis. Educ Psychol Meas 20(1):141–151

Kaiser HF (1970) A second generation little jiffy. Psychometrika 35(4):401–415

Kim J-O, Mueller CW (1978) Factor analysis: statistical methods and practical issues, vol 14, Sage

Klatt DH (1980) Software for a cascade/parallel formant synthesizer. J Acoust Soc Am 67(3):971–995

Kline RB (2013) Exploratory and confirmatory factor analysis. In: Petscher Y, Schatschneider C (eds) Applied quantitative analysis in the social sciences, pp 171–207

Kraft V, Portele T (1995) Quality evaluation of 5 german speech synthesis systems. Acta Acust 3(4):351–365

MacCallum RC, Widaman KF, Zhang S, Hong S (1999) Sample size in factor analysis. Psychol Methods 4(1):84–89

Mayo C, Clark RA, King S (2005) Multidimensional scaling of listener responses to synthetic speech. In: Proceedings of the 6th Annual Conference of the International Speech Communication Association (Interspeech), pp 1725–1728

Morris J (1995) Observations: SAM: the self assessment manikin, an effecient cross-cultural measurement of emotional response. J Advert Res 35(6):63–68

Mulaik SA (2009) Foundations of factor analysis, 2nd edn. CRC Press, Boca Raton

Mullennix JW, Stern SE, Wilson SJ, Dyson C-L (2003) Social perception of male and female computer synthesized speech. Comput Hum Behav 19(4):407–424

Norrenbrock C et al (2015) Quality prediction of synthesized speech based on perceptual quality dimensions. Speech Commun 66:17–35

Oort FJ (2005) Using structural equation modeling to detect response shifts and true change. Qual Life Res 14(3):587–598

Pett MA, Lackey NR, Sullivan JJ (2003) Making sense of factor analysis: the use of factor analysis for instrument development in health care research, Sage

Rosseel Y (2012) lavaan: an R package for structural equation modeling. J Stat Softw 48(2):1–36

Sass D (2011) Testing measurement invariance and comparing latent factor means within a confirmatory factor analysis framework. J Psychoeduc Assess 29(4):347–363

Schwartz CE, Sprangers MA (1999) Methodological approaches for assessing response shift in longitudinal health-related quality-of-life research. Soc Sci Med 48(11):1531–1548

Syrdal AK, Kim Y-J (2008) Dialog speech acts and prosody: considerations for TTS. In: Proceedings of Speech Prosody. pp 661–665

Tabachnick BG, Fidell LS (2014) Using multivariate statistics. Allyn and Bacon, Boston

Taherdoost H, Sahibuddin S, Jalaliyoon N (2014) Exploratory factor analysis; concepts and theory. In: Advances in Applied and Pure Mathematics, pp 15–17

Thompson B (2004) Exploratory and confirmatory factor analysis: Understanding concepts and applications. American Psychological Association

Thurstone LL (1947) Multiple factor analysis: A Development and Expansion of the Vectors of Mind. University of Chicago Press, Chicago, p 535

Tokuda K, Zen H, Black AW (2002) An HMM-based speech synthesis system applied to English. In: Proceedings of IEEE Workshop on Speech Synthesis. IEEE, pp 227–230. doi:10.1109/WSS.2002.1224415

Tseng A, Bansal R, Liu J, Gerber AJ, Goh S, Posner J, Colibazzi T, Algermissen M, Chiang I-C, Russell JA et al (2014) Using the circumplex model of affect to study valence and arousal ratings of emotional faces by children and adults with autism spectrum disorders. J Autism Dev Disord 44(6):1332–1346

Tucker LR, MacCallum RC (2016) Exploratory factor analysis, Unpublished manuscript, Ohio State University, Columbus

Vandenberg RJ, Lance CE (2000) A review and synthesis of the measurement invariance literature: suggestions, practices, and recommendations for organizational research. Organ Res Methods 3(1):4–70

Viswanathan M, Viswanathan M (2005) Measuring speech quality for text-to-speech systems: development and assessment of a modified mean opinion score (MOS) scale. Comput Speech Lang 19(1):55–83

Wicherts JM, Dolan CV (2010) Measurement invariance in confirmatory factor analysis: an illustration using IQ test performance of minorities. Educ Meas Issues Pract 29(3):39–47