Statistical Evaluation of Reliability of Large Scale Listening Tests

Daniel Tihelka[1] and Jan Romportl[2]
[1]*Depart. of Cybernetics, Faculty of Applied Sciences, University of West Bohemia, Czech Republic*
[2]*SpeechTech, s.r.o., Czech Republic*
dtihelka@kky.zcu.cz, jan.romportl@speechtech.cz

Abstract

The present paper deals with the evaluation of large–scale listening tests and with the detection of unaccountable or unreliable answers for each listener. The iterative maximum likelihood estimation scheme is proposed and its abilities are demonstrated and discussed on data collected from a large-scale listening test which was carried out with the aim to collect reference material capturing human perception of similarity of suprasegmental speech units.

1. Introduction

In the field of text-to-speech synthesis, listening tests are still essentially the only means of synthetic speech quality evaluation. The wide range of tests, evaluation intelligibility or naturalness (or overall quality) are either formally standardised or de-facto considered as standards [1, 2, 3]. However, the present paper will focus on the general aspects related to purpose-specific listening tests, designed with the aim to collect data on which the estimate of humans' behaviour can be established [4, 5], as there is no general nor standardised methodology proposed for such kind of tests. Although the paper does not deliver definite conception or "standard", since wider consensus is necessary, it raises some questions which, we believe, should be taken into consideration. They are mostly connected with the reliability of data collected from a wider range of test participants (listeners), and with the possibilities of detecting unaccountable answers caused either non-intentionally by task ambiguity, misunderstanding or by mistake, or worse, by participants' sloppiness or even cheating. The problems are illustrated on a real listening test aimed to qualify the initial concept of how the dissimilarity of the different spoken variants of prosodic words is perceived by humans, described in detail in [5].

2. Similarity Perception Listening Test

When attempting to find a measure (on the acoustics signal) of two prosodic patterns which would provide a sufficiently solid estimate of how the similarity of the patterns is perceived by humans, we need to have a reliable model of such human perception. Due to the variability in both human speech production and perception, it is preferable to have a reliable estimate of unmeasurable true reality, i.e. how similarly the patterns are perceived by humans, and to relate measures on signal to that estimate (as described in [5]), although the building of such an estimate needs to be based on a robust dataset. The opposite approach (i.e. defining a measure which determines an ordering of prosodic patterns with respect to their perceived similarity estimated by the measure, and then to let listeners check how the ordering matches their opinions regarding the similarity) is much less effective due to the need of re-checking the reliability of the similarity estimate every time a new measure is designed, even if each particular test could, theoretically, be smaller in scale.

To obtain the estimate mentioned, we have designed a large–scale listening test to collect the dataset of how the dissimilarity/similarity of prosodic patterns, embodied by two spoken variants of the same word[1], is perceived by humans. The test consisted of 780 stimuli created by pairwise–combining all variants for each of the 17 words. The signals of the words were obtained from a female corpus recorded for our TTS system ARTIC [6], each word cut on boundaries given by automatic segmentation, manually checked and faded in and out to suppress the influence of surrounding words. The listening tests themselves were organised on the client–server basis, using specially developed web application, and due to quite a large size of the test, the participation

[1]To be precise, *prosodic words* were used in the test. However, from the point of view of this paper, the two terms may arbitrarily be exchanged, and thus, for simplicity the term *word* will be used.

(and correct finishing) was financially rewarded. Students from all faculties of our university were addressed by show–cards and students' information web, so virtually any student was able to take part in the test. Nevertheless, only 63 participants finished the tests (accounts of others were closed in the test application when they were not active for several days).

Before each participant started the test, he/she had been familiarised in detail with the purposes of the tests [5] as well as with his/her task – to judge the feeling of the dissimilarity[2] of two versions of a word, using one of the following levels

- *clearly dissimilar* – clear after the very first listening,
- *dissimilar* – quite close but still recognisably not the same,
- *quite similar/indistinguishable* – being very close even if differing after careful listening, or not recognisable at all.

For each word pair, the dissimilarity was requested to be evaluated on all of the following categories (resulting in 4 values)

- *timing* – meaning how much the words differ in rhythm, if there are differences (and how large) in shortening or lengthening in different parts of the words,
- *intonation* – the level of differences in the melody of words, various intonation peaks or valleys within words or melody tendencies spanning the words (e.g. rising contra falling), or differences in stress and/or accent (but not differences in overall pitch level),
- *voice colour* + *pitch level* – the level of difference in voice colour and/or overall pitch level as such, regardless of other aspects (especially melody),
- *overall feeling* – the level of difference of words as such, on all the qualitative levels on which the acoustics is perceived and a difference can be felt. It should capture a "factor X" (e.g. pronunciation) causing words to sound different while all the previous factors do not differ very much.

8 examples of exemplary evaluations for all the aspects were presented for listening, aiming to delimit a notion of individual categories. The participants were, however, urged to rely on their own feeling, as following instructions too strictly is likely to devalue the "objectiveness" expected to emerge from the range of subjective judgement.The categories were chosen on the basis of the listening test data analysis before the test was started, as well as our intuitive reasoning, and they are aimed at the study of the dissimilarity relations/prominences of distinctive prosodic constituents. Moreover, the variants of a prosodic word in all test queries were presented in the order AB to one half of the listeners, and in order BA to the other half (the ordering was selected at random), to study dissimilarity asymmetry [5], and to minimise its effect when all data are used for further processing. Nevertheless, such analyses are beyond the scope of this paper.

One of the check mechanisms employed to help with participants' reliability detection was the repetition of 15 queries randomly placed through the test, aiming to expose participants not self-consistent (i.e. using *clearly dissimilar* and *quite similar* evaluations for the same word pair). In addition, we have identified our expectations on 103 queries selected randomly, but instead of choosing one of defined dissimilarity levels, a likelihood of evaluation was assigned to each level, as there is no guarantee that our opinions, although possibly more qualified, represent the truth (why we did so will be more clear from the discussion at the end of Section 3). Although the evaluations not passing either of check mechanisms can be penalised or excluded directly, it leads, however, to the reduction of responses set and thus of statistical relevance of the dataset. And even if the particular participants were asked to review the whole test (which should, due to the sheer scale of the test, be paid again, making the test more expensive), we are still not able to conclude anything about the reliability of the answers in any other queries except those repeated and reference – there may remain many unaccountable answers through the test unknown.

The need to make the evaluations as robust as possible (the attempt to find acoustic correlates to an estimate of (dis)similarity based on unreliable evaluations is simply not supposed to be very successful), requires to detect unaccountable answers all through the test and make the particular participants review them. It becomes even more important when the reliability of cross-participant agreement computed by means of Fleiss' kappa [7] is only 0.21, which does reject the null hypothesis that observed agreement is accidental on significance level 0.05, but does not make the agreement strong enough to be used as a solid basis for finding definite acoustic correlates[3].

[2] The reasons why not similarity, but dissimilarity was evaluated are discussed in [5]. For the purposes of this paper it does not matter if the statistical processing described further is applied to data representing similarity, dissimilarity, phrase break occurrences or any other subject to which a listening test is focused on.

[3] Although one of interpretations may be that there is no paradigm of speech similarity perception on which humans would agreed, we expect that such a small score is caused by the large scale of the test together with the fact that participation was rewarded no matter the result. For the revision discussed in Section 4, we prepare motivation scheme significantly advantaging participants on the basis the pass of

3. Algorithm for Statistical Evaluation

If we have no a priori notion of the nature of the evaluated listening test, or we deliberately intend not to have one, the intuitive principle of majority is usually used to determine the most prominent answer for a test query (or relative prominence of all answers can be taken into account, if preferable). For further reading, let us suppose that each query can have assigned one $k \in K$ answer by each listener, where $k = 1, \ldots, l$ is an indexing of l possible answers (evaluations) which a listener is allowed to assign to each query. Imagine now the case when for a given test query evaluated by m participants, the relative prominence is $r_k \approx 1/m,\ \forall k \in K$ or more generally, the relative prominence of some answers is $r_k \approx r_x,\ k \in \mathcal{K},\ x \in \mathcal{K},\ \mathcal{K} \subset K$ and $r_k > r_y,\ y \in (K - \mathcal{K})$. Although we may still be able to determine the answer with prominence slightly higher than the others, or we can choose randomly if several prominences are equally the best, how can we be sure that the chosen answer is the right one? What if there are a significant number of participants (not majority, though) who tended to choose an unaccountable answer for the query (no matter if due to task ambiguity, misunderstanding, or by mistake), causing the increase in the prominence of that answer?

Therefore, to reduce the impact of extraordinarily biased answers, the adjusted *maximum likelihood estimation* has been employed for the first time in [4] to process listening test responses focused on the phrase boundary detection problem. Since in the paper the algorithm is described for dichotomy only (the phrase boundary may be perceived or not), and $dim(K) = 3$ in our case, let us describe the algorithm in the very general case.

Formally, let the sequence of listening tests queries X represent the (discrete) states of a random process defined as

$$X = \{X_t : t \in T\} \tag{1}$$

where $T = \{1, \ldots, n\}$ is an ordinal numbering of n queries in listening tests, and each X_t is a random variable holding exactly one of the K possible answers for each query. Let us note that X_t are not observable, and in the case of listening tests they are also independent $\forall t$ (queries are usually arbitrarily ordered, not related to each other).

Be then the m test participants numbered by the set $J = \{1, \ldots, m\}$. We can now define m (discrete) random processes $O^{(1)}, \ldots, O^{(m)}$ representing the participants' responses (observations of the random process X) such that

$$O^{(j)} = \left\{ O_t^{(j)} : t \in T \right\} \tag{2}$$

where $O_t^{(j)}$ are random variables, holding also exactly one of the K possible answers assigned by a j^{th} participant for t^{th} query.

The aim can now be formulated as follows: knowing the observations $O^{(1)}, \ldots, O^{(m)}$ (responses for particular queries in the listening tests from each participant), we want to estimate the hidden sequence of states of the process X which best satisfies the given observations (standard maximum likelihood estimation)

$$X^\star = \arg\max_X P(X|O) \tag{3}$$

As we do not know (or intentionally pretend not to know) anything about the relevance of evaluations from individual users, $\forall o \in K$ and $\forall x \in K$ we can define the probability

$$P(O = o|X = x) = p_{o,x} \tag{4}$$

$$\sum_{\forall o} p_{o,x} = 1 \tag{5}$$

expressing the expected likelihood of any participant's ability to correctly identify the hidden state of process $o = x$, and the likelihood of miss if $o \neq x$ (all combinations giving confusion matrix). Although the probabilities are initialised equally for all participants (based on the nature of the tests), the iterative training process described further is going to discriminate them for each participant independently (based on the ability to agree/disagree with the evaluation of other participants).

We can also define the probability of individual states of the process X

$$P(X = x) = p_x \tag{6}$$

$$\sum_{\forall x} p_x = 1 \tag{7}$$

whereby it is desirable, as discussed in [4], not to make any strong a priori assumptions about the process and its distribution. By obeying the principle of Occam's razor, let us presuppose that X is a stationary process with the uniform probability distribution $p_x = p_y \forall x \in K,$ and $y \in K$.

Given the definitions (3)–(7) and the assumption about the mutual independence of X_t, *naïve Bayes model* [8] can be used to estimate $P(X_t|O)$ for each t independently

$$P(X_t = x|O_t) = \frac{\prod_{j \in J} P(O_t^{(j)}|X_t = x) P(X_t = x)}{P(O_t)} \tag{8}$$

check mechanisms employed and the agreement with the most probable estimates described in Section 3.

where $P(O_t)$ can be ignored (and usually is), since it is not a function of x and thus it does not affect the relative values of probabilities for any x. We, however, "normalise" the values given by (8) so that $\forall x, \forall t$ the proportions among $P(X_t = x|O_t)$ would stay preserved, but $\sum_{x \in K} P(X_t = x|O_t) = 1$; the normalised values are then used as resulting $P(X_t = x|O_t)$.

To maximise $P(X|O)$, we employed the iterative procedure which can be considered to be EM algorithm simplified to suit the needs of the described problem. In each step, Equation (8) is computed using the current estimates of Equation (4) and (6). Informally said, on the basis of the likelihood for the individual participant's ability to identify or miss the hidden state of process X, the probabilities of individual states are estimated $\forall t$. As in the very first step, equal likelihoods are used for all participants, the most probable state is the one with the highest inter-participant agreement (the highest number of $P(O_t = o|X_t = x)$ where $x = o$, are multiplied). Then the likelihoods (4) and (6) are recalculated for each participant and state independently $\forall j \in J, \forall o \in K$ and $\forall x \in K$

$$P(O^{(j)} = o|X = x) = \frac{\sum_{t \in T} P(X_t = x|O_t = o)}{\sum_{t \in T} P(X_t = x|O_t)} \tag{9}$$

$$P(X = x) = \frac{\sum_{t \in T} P(X_t = x|O_t)}{\dim(T)} \tag{10}$$

obtaining new likelihoods of participant identification abilities as well as process state probabilities, now based on given observations. In this step, the corresponding probabilities of miss are increased (and probabilities of correct identification are lowered) for all those participants whose responses differ from the most probable states estimated by Equation (8), and the other way round. Let us note that the described process converges to a local maximum, and hence the initial parameters are recommended to be chosen reasonably and perturbed in more experiments.

Although such treatment allows us to detect consistent misses (answering $o \in K$ in majority of cases when most of others answered $p \in K,\ o \neq p$) for each participant, manifested in higher probability of miss in Equation (4), the results must still be interpreted with care – all the probabilities are estimated according to cross-participant agreement, not knowing anything about a "real" state of process X, as we defined it.

In reality, we must always presume that a number of participants will tend to place unaccountable answers, and the question now is how to detect such answers. If we have reference evaluations of queries subset $\mathcal{T} \subset T$ at our disposal, we can force them in each iteration to $P(X_t|O_t), \forall t \in \mathcal{T}$ instead of original values computed by Equation (8). In this way, when computing Equation (4) the probabilities of correct identification will be lowered for all users with evaluations differing from the reference. Although it relaxes our intention not to make any strong a priori assumptions about the process X, the intention was already relaxed by the definition of the expectations in check point queries. Still, we must be careful. The more knowledge about process X we force into Equation (8), the more the process will be pushed to that knowledge – it negates the sense of an "objectiveness" emerging a from wide range of subjective opinions.

4. Evaluation of Dissimilarity Perception

To evaluate the results of listening tests, described in Section 2, by means of the proposed algorithm, we initialised the confusion matrix of all participants to

$$P(O^{(j)}|X) = \begin{pmatrix} 0.5 & 0.3 & 0.15 \\ 0.35 & 0.4 & 0.35 \\ 0.15 & 0.3 & 0.5 \end{pmatrix} \qquad \forall j \in J \tag{11}$$

being quite benevolent in exchanging *quite similar/undistinguable* with *dissimilar* and *clearly dissimilar* with *dissimilar* (and the other way round). As we really know nothing about dissimilarities distribution in the test set (all word pairs have the same initial probability of being clearly dissimilar, dissimilar or similar/indistinguishable), the probabilities $P(X)$ were initialised to

$$P(X = x) = 1/3 \qquad \forall x \in K \tag{12}$$

The evaluations for each aspect were processed separately, to avoid the interfusing of confusion matrixes $P(O^{(j)}|X)$ – each listener can rather consistently make some kind of characteristic errors which may, however, differ for individual aspects.

As the first part, the summary of differences between reference evaluations of the 103 queries and $X^\star$ obtained by majority agreement and the proposed maximum likelihood estimation is presented. In Table 1, the number of queries matching the reference answer $\overline{x}$ (the one with the highest likelihood assigned by us) are shown when determined by simple majority agreement, by $X^\star$ from the proposed estimation scheme and by $X^\star$ the proposed estimation with the reference likelihoods forced into Equation (8), as described at the end of Section 3.

It can be seen that even the estimation *not taking* reference data into account (i.e. nothing is known about an expected evaluation at all) is usually able to provide a better match to the expectation $x^\star = \overline{x}$, while the number of $x^\star = \overline{x}$ in cases when a majority of participants have chosen $o \neq \overline{x}$ is referred in Table 1 as $\uparrow x$. This

Table 1: The comparison of results obtained by *majority* agreement, the proposed *estimation* method and the estimation method with *reference* evaluations forced into Equation (8). Columns represent how many answers match, how many differ by one dissimilarity evaluation level and how many by two levels.

overall dissim.	match			1 level	2 levels
majority	51	$1\downarrow$	$3\downarrow\downarrow$	50	2
estimation	59	$\uparrow 9$		41	3
estim+refer.	61	$\uparrow 5$	$\uparrow\uparrow 13$	40	2
tempo dissim.	match			1 level	2 levels
majority	74	$9\downarrow$	$10\downarrow\downarrow$	25	4
estimation	74	$\uparrow 9$		28	1
estim+refer.	76	$\uparrow 3$	$\uparrow\uparrow 12$	25	2
melody dissim.	match			1 level	2 levels
majority	51	$14\downarrow$	$14\downarrow\downarrow$	50	2
estimation	71	$\uparrow 34$		32	0
estim+refer.	72	$\uparrow 1$	$\uparrow\uparrow 35$	31	0
colour/pitch	match			1 level	2 levels
majority	62	$8\downarrow$	$11\downarrow\downarrow$	39	2
estimation	66	$\uparrow 12$		36	1
estim+refer.	66	$\uparrow 3$	$\uparrow\uparrow 15$	35	2

is due to the fact that a sufficient number of listeners with higher reliability $P(O = o|X = x)$ of determining a dissimilarity level $o = x^\star$, $x = x^\star$ (from the point of view of the algorithm) have really chosen that level, even when the majority of listeners with lower reliability have chosen a different answer. On the other hand, one must realise that the same principle may lead to the choice of answer $x^\star \neq \overline{x}$, even when a majority of listeners have really chosen $o = \overline{x}$ (in Table 1 referred to as $x\downarrow$) – it is a natural cause of the fact that the estimation process relies only on given cross-listener agreements, as discussed in Section 3. When the probabilities of expectations assigned to reference queries were enforced into the estimation procedure, the number of $x^\star$ equal to $\overline{x}$ was, es expected, even higher. The cause, again, is that the likelihood of answering $o = \overline{x}$ was slightly increased for participants answering in that way. The answers of such participants were, therefore, able to outweigh the majority answers, if they differed (referred as $\uparrow x$ when compared to estimation unaffected by reference data, and as $\uparrow\uparrow x$ when compared to majority). However, there are still queries with $x^\star \neq \overline{x}$. The reason, once again, is that there are participants with high reliability of their answers (possibly also increased by answering $\overline{x}$ in other queries), whereby their answers differ from $\overline{x}$. Despite the increase in match, there are still queries with $x^\star$ different by two levels (e.g. $\overline{x}$ = *quite similar* and $x^\star$ = *clearly different*). Since they again occurred by the agreement of highly reliable participants (from the point of view of the algorithm), the evaluations of such participants must be further analysed in depth. Let us also stress that nothing more can be concluded about the other (non–reference) queries, except the assumption that $X^\star$ determined for those queries is likely to be closer to what we feel to be the expected evaluation.

Contrary to [4], where the aim was to obtain $X^\star$ best satisfying the given evaluations, in our work the concrete dissimilarity evaluations of each individual participant for each individual query is used to build a dissimilarity matrix in [5]. However, as the scale of the test is quite large, we can expect the occurrences of unaccountable evaluations which may unpredictably bias the estimates of perceived dissimilarities and thus to significantly distort the conclusions. Therefore, the issue now is to somehow detect the queries with the unaccountable evaluations individually for each participant, and make these participants review their answers on such queries (providing them with more examples about consistent misevaluations obtained from the reference data) – the revaluation of the whole test set by all participants is virtually impossible, due to the sheer scale of the test. The simplest approach may be to make all listeners reassess only their queries differing from $X^\star$. This is neither ideal (still results in huge number of queries to reassess) nor required ($X^\star$ evaluations can directly be used for all listeners). Instead, it seems to be feasible to take advantage of the values in $P(O|X)$, where the answers of listeners with a high miss likelihood might be expected as not very reliable – even the increase in match (or in mismatch) to reference answers is reflected in those values. Let us, therefore, to review all queries t for participant j with likelihood of miss

$$P(O^{(j)} = o|X = x^\star) > E(P(O = o|X = x^\star)) + \\ + k\sigma(P(O = o|X = x^\star)), \\ o \neq x^\star \tag{13}$$

where constant k allows us to make the requirement on miss more strict, or to alleviate it.

In Table 2, the results for $k = 0$ and $k = 1$ are presented. We have decided to choose $k = 1$ due to reasonable average number of review requests, while the participants with the highest likelihood of miss are still determined. Unfortunately, the reviews were not finished at the time of writing, so we cannot present the shift of the differences of reviewed evaluations to reference data in contrast to results in Table 1, the shift of Fleiss kappa

Table 2: The comparison of the required number of queries to review for different k in Equation (13). The columns represent the *average*, *minimum* and *maximum* number of requests per participant; *partic.no.* is the number of participants who must review at least one query.

overall dissim.	avg.	min.	max.	partic.no.
$k = 0$	230.0	72	513	62
$k = 1$	120.1	20	403	37
tempo dissim.	avg.	min.	max.	partic.no.
$k = 0$	179.5	47	338	61
$k = 1$	90.4	57	321	33
melody dissim.	avg.	min.	max.	partic.no.
$k = 0$	288.7	120	487	63
$k = 1$	130.5	22	447	41
colour/pitch	avg.	min.	max.	partic.no.
$k = 0$	233.1	56	376	62
$k = 1$	102.6	75	363	33

value, nor the change of average $P(O|X)$ and its standard deviation or variance.

5. Conclusion

As the main aim of the paper we have proposed an iterative algorithm estimating the most likely evaluations of listening tests queries, with rigorous maximum likelihood estimation process as the underlying principle. Although there may exist even more powerful means of ambiguous or even "untrustworthy" data analysis (despite requiring some additional a priori information), we have attempted to show that the algorithm may be a well-founded option to the evaluation based on the expectation that majority agreement among listeners represents an "objective" truth emerging on the examined data. On top of that, the algorithm also provides some means of determining unaccountable evaluations for further review. We also plan to use the agreement to the most likely evaluation estimated as one of motivation factors determining the amount of financial reward in the follow-up revision process.

6. Acknowledgement

This research is supported by the Grant Agency of the Czech Republic, project no. GACR 102/06/P205, and by Ministry of Education of the Czech Republic, project no. 2C06020. Special thanks must also go to our colleague Jan Zelinka for his valuable consultations on the proposed algorithm.

7. References

[1] ITU-T Recomendation P.800: Methods for objective and subjective assessment of quality, 1996

[2] C. Benoit, M. Grice, V. Hazan, "The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using Semantically Unpredictable Sentences", *Speech Communication*, 18, pp. 381–392, 1996.

[3] D. Tihelka, J. Matoušek, "The Design of Czech Language Formal Listening Tests for the Evaluation of TTS Systems", in Proc. of LREC, vol. VI, pp. 1099–2002. Lisbon, Portugal, 2004.

[4] Romportl, J., "Statistical Evaluation of Prosodic Phrases in the Czech Language", in Proc. of Speech Prosody, Campinas, Brasil 2008.

[5] Tihelka, D. "Towards Automatic Measure of Similarity for Use in Unit Selection", Submitted to ICSP 2008. Beijing, China, 2008.

[6] Matoušek, J., Romportl, J., Tihelka, D., and Tychtl, Z. "Recent Improvements on ARTIC: Czech Text-to-Speech System", in Proc. of ICSLP. vol. III, pp. 1933–1936. Jeju, Korea, 2004.

[7] J.L. Fleiss, "Measuring nominal scale agreement among many raters", in Psychological Bulletin, vol. 76, no. 5, pp. 378–382. 1971.

[8] D. Barber, "Learning from data 1: Naive Bayes". Online: http://axiom.anu.edu.au/∼daa/courses/GSAC6017/naivebayes.pdf, accessed on 6 June 2008.