Robust methodology for TTS enhancement evaluation*

Daniel Tihelka, Martin Grůber, and Zdeněk Hanzlíček

University of West Bohemia, Faculty of Applied Sciences, Dept. of Cybernetics
Univerzitni 8, 306 14 Plzeň, Czech Republic
{dtihelka,gruber,zhanzlic}@kky.zcu.cz

Abstract. The paper points to problematic and usually neglected aspects of using listening tests for TTS evaluation. It shows that simple random selection of phrases to be listened to may not cover those cases which are relevant to the evaluated TTS system. Also, it shows that a reliable phrase set cannot be chosen without a deeper knowledge of the distribution of differences in synthetic speech, which are obtained by comparing the output generated by an evaluated TTS system to what stands as a baseline system. Having such knowledge, the method able to evaluate the reliability of listening tests, as related to the estimation of possible invalidity of listening results-derived conclusion, is proposed here and demonstrated on real examples.

Keywords: speech synthesis, TTS evaluation, listening tests, statistical reliability

1 Introduction

During the development, enhancement, or experiments with text-to-speech (TTS) system, researchers usually face the problem of reliable evaluation of the new or enhanced system. Contrary to speech recognition (ASR, LVCSR) [**?**,**?**], which can be evaluated by using an input signal to be recognised accompanied by the text transcript of the signal expected to be recognised, and the use of mathematical methods of text difference evaluation, the evaluation of a TTS system must still rely on (rather a larger number of) subjective responses of listeners evaluating the naturalness, or comparing two versions of a TTS system. Unfortunately, the comparison of TTS outputs on the signal level, although being mathematically rigorous, does not correspond very much to the listeners' perception [**?**,**?**].

In addition, there is a problem with the absence of a unified methodology of comparing various TTS systems. With the exception of the traditional Blizzard Challenge evaluation event [**?**,**?**], there is no common public database (or a set

* The research has been supported by the Technology Agency of the Czech Republic, project No. TA01030476 and by the European Regional Development Fund (ERDF), project "New Technologies for Information Society" (NTIS), European Centre of Excellence, ED1.1.00/02.0090.

of language-specific databases) from which synthetic speech could be created and compared to speech generated by other synthesizers. The existence of such, within TTS community generally agreed, database would allow the robust and reliable comparison of TTS systems with results mutually comparable across individual systems, whether evaluating the enhancement of a particular TTS engine, or comparing one engine to another. Without this, the TTS researchers will continue to face the problem of how to validly interpret non-self-repeatable private-data-collected evaluations like "on our data, the performance was increased significantly", which we were faced with e.g. in [**?**].

However, the situation in the field of TTS evaluation is even worse. The usual procedure is to randomly select [**?**,**?**,**?**,**?**,**?**] a number of sentences, synthesize them by various synthesizers (or by various versions of the same synthesizer) being compared, and let the listeners evaluate individual outputs independently (using the MOS test), or confront one with the other (using the CCR test). The overlooked side of such an approach is that it may not reveal very much about the enhancement or the lowering of the quality, since the significant cases (see the definitions of δ for what they mean) that may show different results have not even been tested.

Therefore, the present paper aims to offer a rigorous methodology which enables lining up and quantifying the possibility that the results drawn from listeners' evaluation may not be entirely trustworthy. We are convinced that this information should be included in each listening test-based TTS evaluation.

2 Reliability of listening tests

When experimenting with enhancements of a TTS system, it is usually a rather small part of the whole system which is actually changed. Therefore, when synthesizing a set of test stimuli by the baseline and the enhanced version, some parts of the generated speech and even whole phrases will remain identical for both versions.

However, to be able to evaluate results, only a reasonable number of stimuli is generated for the listening test, usually ranging from 10 to 30. Otherwise, i.e. when having a larger number of listening stimuli, there is a risk of not finding a sufficient number of listeners willing to participate in the test, or they will not carry out the test as carefully as they should.

2.1 Suitability of phrases to be evaluated

Let us define a difference function $\delta(a, b)$ determining how much two variants a and b of the same phrase (generated by different synthesizers or synthesizer versions) differ from the point of view of the material used to build the phrases:

$$\delta(a, b) \in \langle 0, 1 \rangle$$

with boundary value $\delta(a, b) = 0 \Leftrightarrow a = b$ and $\delta(a, b) = 1 \Leftrightarrow a \cap b = \emptyset$. For example, in the case of unit selection TTS, the $\delta(a, b) = 1$ is the case where no

unit candidate occurring in a is also used in the same position in b. Nevertheless, it is not strictly necessary to limit the range in any way; we have simply used it for higher readability.

Having the difference function, we can build its probability mass function $P(\delta)$ (shown in Figure 1) and distribution function $F(\delta)$. Note that we need to work with probability space, since the set of phrases to be synthesized, and to be possibly listened to, is countable but not finite[1]. And now it is simple to compute

$$P(X \geq \delta) = 1 - P(X < \delta)$$

which represents the probability of synthesizing a phrase with the value δ higher than a given value.

The interpretation of this number is now rather straightforward and provides important information usually neglected in the "classic" approach — how large is the expected probability of the occurrence of a synthetic phrase not covered by the listening tests, i.e. the probability of the occurrence of a phrase for which the result of the listening test is not valid. Or alternatively, what is the probability of the occurrence of the worst possible case covered by the tests (the δ of phrases used for listening), while one can expect (or test) that all of the better cases (lower δ) will not sound worse, simply due to their higher similarity to the version of TTS system being compared to[2].

2.2 Real example demonstration

To illustrate this on some real examples, let us use the statistics obtained during the research into our TTS system optimization (described in [**?**]) as well as a "dummy" experiment in which we changed a small part of concatenation cost computation:

data size reduction, experiment 1. We have removed some phrases from the full corpus to reduce it to approximately 66% of the original size, which is a size acceptable for the joining with screen-reader programs used by blind or semi-blind people on their home PCs. In this experiment, the synthesis is actually forced to use different unit candidates than those used by the baseline, since they are absent from the reduced dataset.

data size reduction, experiment 2. It is basically the same as the previous experiment, except that the number of phrases removed was much larger – the reduced data size is approximately 17% of the original data. Thus, the ratio of different candidates is much higher here.

[1] In general case. For pragmatic reasons, however, we limit the length of phrases for listening tests to a certain length and to natural sentences only.

[2] Of course, even a small change in a single place may decrease the quality of the TTS speech, but it is still more probable that if this single change makes the speech worse, more changes will make it even worse. Or to put it differently, it is very unlikely that when synthetic speech with higher δ is evaluated consistently better, synthetic speech with lower δ will sound consistently worse.

unit selection feature change. It is a kind of artificial example of unit selection algorithm tuning, where we slightly changed concatenation cost computation – instead of Euclidean distance between 12 MFCC vectors, the average absolute difference of first 2 MFCC is computed. Contrary to the previous experiments, the number of units remain the same and the differences in the output (if there are any) are caused by the different behaviour of the unit selection algorithm.

The outputs of the modified system were always compared to the baseline of our TTS. In the illustration, the following two schemes of δ computation were used:

1. the difference of unit candidates. For a and b variants of a phrase consisting of N units with candidates in sequence $a_1, a_2, \ldots, a_N$ and $b_1, b_2, \ldots, b_N$ respectively, the difference is computes as

$$\delta^k(a,b) = \frac{\sum_{i=1}^{N} \|a_i, b_i\|}{N}$$

   where $\|a_i, b_i\| = 0 \Leftrightarrow a_i = b_i$ and $\|a_i, b_i\| = 1$ otherwise.
2. the difference in the number of concatenation points being defined for the same sequences a and b as

$$\delta^l(a,b) = \frac{|\sum_{i=1}^{N-1} \mathcal{D}(a_i, a_{i+1}) - \mathcal{D}(b_i, b_{i+1})|}{N-1}$$

   where $\mathcal{D}(a_i, a_{i+1}) = 0 \Leftrightarrow$ there is no discontinuous concatenation point between candidates a_i and a_{i+1}, meaning that a_i is the natural and immediate predecessor of a_{i+1} in natural speech, i.e. $next(a_i) = a_{i+1}$*;* $\mathcal{D}(a_i, a_{i+1}) = 1$ *otherwise.*

The charts in Figure 1 represent the probability mass function $P(\delta)$ and its distribution collected by the synthesis of more than 1 million phrases for all the experiments. Of course, the definition of δ depends on the form of TTS tested and on the expectation what may affect the resulting speech; e.g. there is no straightforward way of unit candidates comparison in the case of HMM-based synthesizer.

Let us now assume the "classic" evaluation test procedure without the knowledge of this statistics. For the purpose of the listening test, let us imagine that 30 randomly selected phrases would be synthesized, which is quite an unusually large number. And naturally, we hope (or claim in the results) that the evaluated phrases represent an typical overall behaviour pattern of the TTS work, or rather that the synthesized versions of the selected phrases differ from the baseline preferably more than less. However, according to the Bernoulli schema

$$P_{<x,y>} = \sum_{i=x}^{y} \binom{y}{i} P^i (1-P)^{y-i}$$

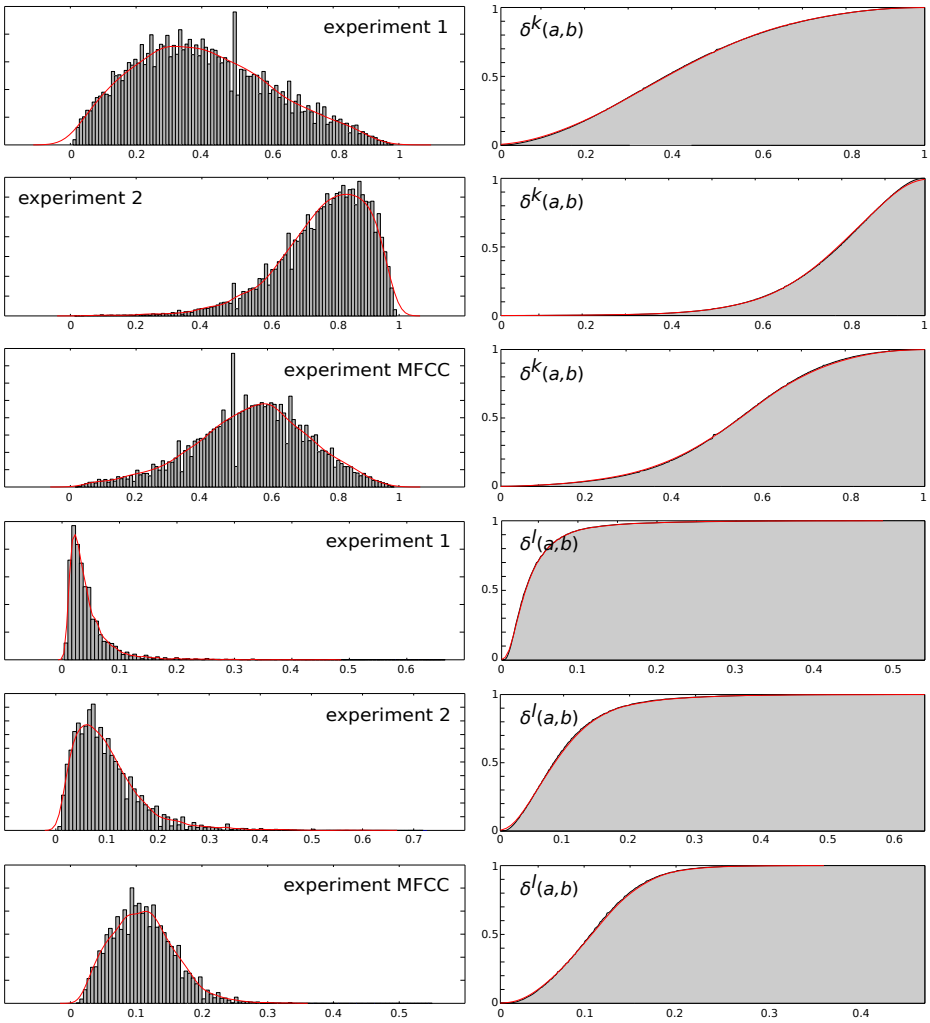

Fig. 1. The visualization of $\delta(a, b)$ probability mass function (sampled to 100 equidistant intervals) and the corresponding distribution function for *speaker 1*. The solid line represents the estimated probability density function (see Section 2.3).

it can be computed that the probability $P = P(X >= \delta)$ for at least 15 of the 30 synthesized phrases (i.e. $x = 16, y = 30$) is not very high, especially for experiments with lower amount of δ changes in the synthesized speech. In Table 1, we have chosen $X >= 0.6$ for δ^k as representing the case with 60% or more of changed units, and $X >= 0.1$ for δ^l meaning 10% or higher change of concatenation points. And, preferably, those would be the cases which the listening tests should focus on, since those cases are more likely to manifest benefits or harms of

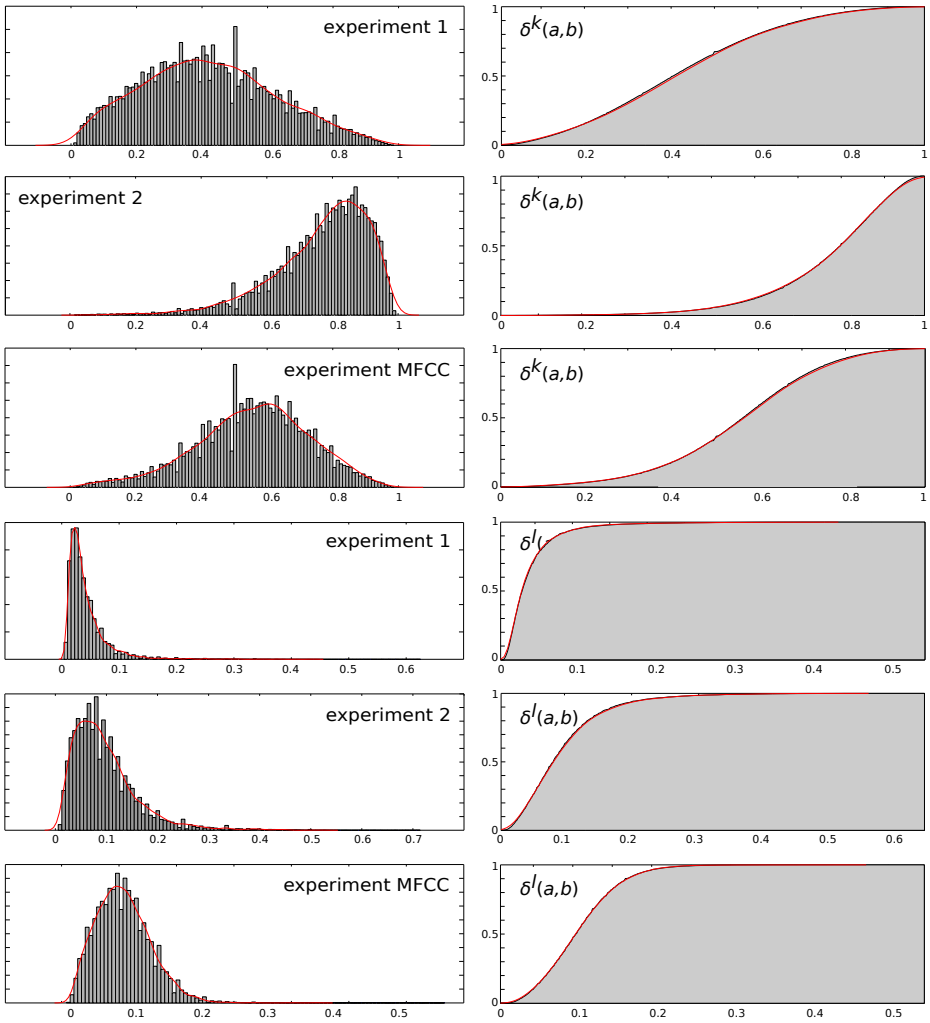

Fig. 2. The visualization of $\delta(a, b)$ probability mass function and the corresponding distribution function for *speaker 2*. The structure is the same as in Figure 1.

the evaluated TTS (for vindication see Footnote 2). We hope that this example clearly shows the need of knowledge (or at least an awareness) of the synthesized data behaviour, as well as the estimation of incorrect evaluation results probability, analogous to significance level in hypothesis testing. Otherwise, the evaluation results (and thus conclusion claims) may not be very representative.

$\delta^k(a,b)$	speaker 1		speaker 2	
	$P(X >= 0.6)$	at least 15×	$P(X >= 0.6)$	at least 15×
experiment 1	0.194	0.00	0.201	0.00
experiment 2	0.887	1.00	0.882	1.00
experiment MFCC	0.409	0.09	0.421	0.11
$\delta^l(a,b)$				
	$P(X >= 0.1)$	at least 15×	$P(X >= 0.1)$	at least 15×
experiment 1	0.075	0.00	0.061	0.00
experiment 2	0.408	0.08	0.386	0.05
experiment MFCC	0.572	0.66	0.545	0.55

Table 1. The illustration of probability of the selection of at least 15 out of 30 phrases with $P(X >= 0.6)$ for $\delta^k(a,b)$ and $P(X >= 0.1)$ for $\delta^l(a,b)$ for all the experiments.

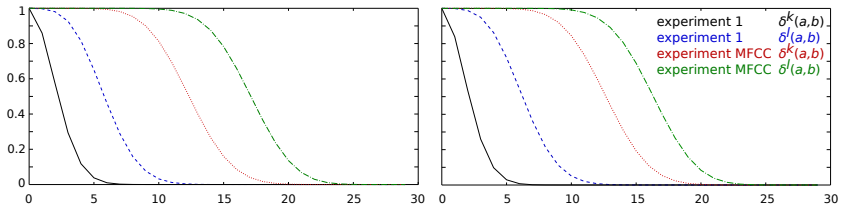

Fig. 3. The visualization P_i for all the cases where at least $i = 1, 2. \ldots, 30$ out of 30 phrases match the required $P(X)$; left for *speaker 1*, right for *speaker 2*.

2.3 Use of the probability with smaller dataset

For the computation, we have directly used the relative occurrence values of δ collected from the 1 million synthesis output. Let us note that the synthesis ran more than 17 hours on MPI-parallelized 32 Intel Xeon X5680 3.33GHz cores, which would take approximately 23 days on the single core. Of course, it is not usually necessary nor meaningful to collect such an extensive amount of data, since standard kernel density estimation technique [**?**] can be used to model the probability distribution function from a much lower δ dataset. To prove it, we have randomly selected $5,000$ δ values, for which we have done this estimate. This case is illustrated in Figure 1 by the solid line and the values of $P_{<16,30>}$ are shown in Table 2.

Simple comparison of the estimated values in Table 2 with data-computed values in Table 1 shows that the estimate is sufficiently precise.

3 Conclusion

We aimed to show that there is a non-negligible (and even computable) possibility of drawing unreliable results from listening test evaluation, when phrases to be evaluated are chosen at random, which is, however, the usual manner of their

$\delta^k(a,b)$	speaker 1		speaker 2	
	$P(X >= 0.6)$	at least 15×	$P(X >= 0.6)$	at least 15×
experiment 1	0.192	0.00	0.204	0.00
experiment 2	0.890	0.99	0.882	1.00
experiment MFCC	0.410	0.09	0.450	0.18
$\delta^l(a,b)$				
	$P(X >= 0.1)$	at least 15×	$P(X >= 0.1)$	at least 15×
experiment 1	0.073	0.00	0.061	0.00
experiment 2	0.431	0.13	0.418	0.10
experiment MFCC	0.590	0.68	0.568	0.65

Table 2. The illustration of the same values as in Table 1, but here computed from distribution functions $F(x)$ get by the kernel density estimation technique.

selection. Moreover, without some knowledge of differences between the outputs from the baseline and the tested TTS, no-one can be sure which of the cases is actually evaluated by the listening tests — frequent cases with slight differences (i.e. $X < \delta$ for smaller δ), or cases with larger changes and thus with higher informative capability (see Footnote 2)?

The point is that the knowledge of $F(X)$ is important for the estimation of the expected listening tests validity. Having the selected set of phrases to be listened to, the minimum, maximum and average δ can be obtained for them and used to compute $P(X \geq \delta)$. This value estimates, among others, the probability of the occurrence of a phrase for which the result of the listening test is not valid (see the interpretation in Section 2.1).

With all this information, the results of various TTS comparisons will gain higher level of reliability and will become more trustworthy for their readers.

The access to computing and storage facilities belonging to the National Grid Infrastructure MetaCentrum was provided under the program LM2010005 “Projects of Large Infrastructure for Research, Development, and Innovations”.