7

Crowdsourced Assessment of Speech Synthesis

Sabine Buchholz[1], Javier Latorre[2], and Kayoko Yanagisawa[2]
[1] *SynapseWork Ltd, UK*
[2] *Toshiba Research Europe Ltd, UK*

7.1 Introduction

Speech synthesis is the artificial generation of human speech, usually by a computer. Since the most common input for a speech synthesizer is text, these systems are often called text-to-speech (TTS) systems. The two main requirements of any TTS system are intelligibility and naturalness. For the first synthesizers created in the 1940s and 1950s, just achieving sufficient intelligibility over normal sentences was a huge task. With the development of computer-based TTS systems and especially after the appearance of data driven systems (waveform-concatenation and parametric-based speech synthesis) in the late 1990s, the quality improved dramatically to the point that intelligibility over normal sentences is mostly a given. At the same time, researchers' focus shifted from intelligibility toward naturalness and those speech features related to it such as similarity to the original speaker and expressiveness (the ability to convey emphasis, emotions, attitudes, irony, character, etc.). However, these features are highly subjective and thus evaluation of such features is more complex and requires more listeners than for intelligibility tests.

This chapter is mainly about crowdsourcing *assessment* of TTS systems, but it also presents some results from attempts to use it for other purposes related to TTS development. Assessment is an important part of the research and development cycle for any type of speech technology. For speech recognition, assessment is usually performed by automatically computing the word error rate (WER) against manually transcribed reference texts. For speech synthesis, some automatic methods have been used that compute the differences between the synthesized

Crowdsourcing for Speech Processing: Applications to Data Collection, Transcription and Assessment, First Edition. Edited by Maxine Eskénazi, Gina-Anne Levow, Helen Meng, Gabriel Parent and David Suendermann.

One way to categorize listening tests is according to the unit of assessment. Segmental tests typically aim to assess performance at the level of individual phones. For example, in the modified rhyme test (MRT; House *et al.* 1963), listeners have to indicate which word they heard in the sample. The word can be embedded in a carrier phrase, but the phrase is always the same, only the one word differs. The words are chosen among minimal pair sets such that mishearing one phone results in a different word. The most common level of TTS assessment nowadays is the sentence level (as most TTS systems' unit of synthesis is still the sentence). Sentence-level listening tests will be discussed in more detail below. Finally, while listening tests involving units beyond the sentence are not new (e.g., ITU 1994), there has been renewed interest for them lately (e.g., Hinterleitner *et al.* 2011; King 2012), due to a desire to move beyond the synthesis of isolated sentences toward the synthesis, and therefore assessment, of longer texts.

Going back to sentence-level assessment, there is a broad division into intelligibility and nonintelligibility listening tests. In intelligibility tests, subjects are asked to type in the words they hear in the synthesized sentence. To prevent listeners from guessing some words by their context, semantically unpredictable sentences (SUS; Benoit and Grice 1996) are often used. In nonintelligibility tests, subjects answer multiple choice questions about the samples. These tests can be further subdivided by the type of multiple choice answers. In mean opinion score tests (MOS; or Absolute category rating, ACR; ITU 1988a), listeners are asked to *rate* one speech sample, typically on a Likert scale (Likert 1932) from 1 to 5. An example is a MOS *naturalness* test, in which listeners have to rate the naturalness of the speech, with 1 being "Completely Unnatural" and 5 being "Completely Natural," (see Fraser and King 2007). It is important to note that MOS only refers to this rating scale of the answer and is largely independent of the exact question asked. For example, ITU (1994) recommends MOS with seven different questions, ranging from "Listening Effort" to "Overall Quality."

In *AB* tests (also called paired or pairwise comparison tests), listeners are typically asked to indicate which of the *two* speech samples has *more* of some property, with the property being defined by the question. For example, in *preference* tests, the question is simply "Which one do you prefer?" For the assessment of expressive synthesis, the question could be "Which one sounds happier (sadder/angrier)?" *AB* tests can be *forced choice*, or allow for a "No preference" option.

In *AB* tests specifically designed to collect data for multidimensional scaling (MDS) analysis, listeners indicate whether two speech samples are the *same* or *different* with respect to some property such as naturalness (see Clark *et al.* 2007).

Orthogonally to the above subdivision of test types, there is another dimension of whether a reference sample is given. For example, while in a normal MOS test only one sample is given, a "degradation" or "differential" MOS (DMOS; ITU 1988b) test in addition provides a reference sample and listeners are asked to rate the sample with respect to this reference. DMOS tests are often used to assess speaker similarity, for example, for speaker conversion or adaptation approaches, and the scale is then labeled "1—Sounds like a totally different person" to "5—Sounds like exactly the same person" (see Fraser and King 2007).

Likewise, while in a normal *AB* test two samples are given, an *ABX* test additionally provides a reference and asks which of the two samples under test sounds more like the reference with respect to the property stated in the question, such as speaker similarity or emotion. Table 7.1 gives an overview of the sentence-level test types discussed.

Table 7.1 An overview of sentence-level listening test types.

Task	Number of samples	Question	Without reference	With reference
Type in	1		Intelligibility e.g., SUS	n/a
Rate	1		MOS e.g., naturalness	DMOS e.g., similarity
Compare	2	Which one more ...?	AB e.g., preference	ABX e.g., expressiveness
Compare	2	Same or different?	For MDS e.g., naturalness	n/a

It is important to note that possibly with the exception of intelligibility tests, all the other types of tests discussed above can only be used to measure *relative* differences between (versions of) TTS systems. In other words, they cannot yield an absolute measure of subjective quality that can be compared across different tests with different systems being compared. For comparison tests, this means that the results are not necessarily transitive. In other words, when comparing three systems A, B and C, it is possible to find that $A > B$, $B > C$ but $C > A$, as have been observed in some of our experiments. This may happen when the differences between systems are not large, for example, $\leq 10\%$, and/or when the tests are not conducted simultaneously by the same group of subjects. In the case of MOS tests, a certain inter-test comparability can be achieved by including natural speech in all the tests to anchor the upper limit of the scale (Taylor 2009), or by always including a benchmark system (Karaiskos *et al.* 2008 onward) but even then, the interpretation of further levels on the scale depends on the systems being compared. As listening tests involve humans and humans cannot be standardized, it is not rare that the same test repeated twice yields different *absolute* results, especially when the group of subjects involved are different. This is relevant for crowdsourcing, as one is almost guaranteed not to get exactly the same set of listeners (workers) on different occasions.

Finally, different test types and instantiations (i.e., test questions) measure different aspects and there is not a single "best" test for TTS assessment. For example, although the final goal of both MOS and *AB* tests is the same, that is, to rank systems according to some subjective criterion, the way they work is different. In MOS tests, the exact score might depend on the previous stimulus the subject has judged (hence randomization is important). However, scores refer to the overall impression over the whole stimulus. Therefore, local differences are usually ignored. In an *AB* test, subjects listen to two versions of the same sentence, which in many cases are very similar to each other. As a result, the overall impression is often the same and what really counts is local differences/errors. This means that paired tests can provide a much finer ranking than MOS tests can. It is not unusual to find significant differences in preference for systems with very similar MOS values. Of course, the disadvantage of *AB* tests is the cost. To evaluate S samples over N systems using *AB* tests $S \cdot N \cdot (N-1)$ stimuli have to be judged, whereas with MOS tests $S \cdot N$ stimuli are sufficient. This disadvantage can partially be counteracted by the abundance of listeners and lower costs of crowdsourced tests.

Sityaev *et al.* (2006) conducted several MOS tests with different test questions and SUS intelligibility tests with three TTS systems. They observed a high correlation between

the "listening effort," "comprehension," and "articulation" MOS tests recommended in ITU (1994), but opposite results for the other tests. While one system scored highest on acceptability, another's voice was liked the most. Also, one system scored the highest in the intelligibility test, but lowest in the MOS naturalness and "overall quality" tests. When comparing different TTS systems, the de facto standard of the Blizzard Challenges—an annual open evaluation of TTS systems—is to at least test intelligibility as well as naturalness.

7.3 Crowdsourcing for TTS: What Worked and What Did Not

7.3.1 Related Work: Crowdsourced Listening Tests

A sizable number of papers have been published in the last 2 years whose main focus is TTS research but which use crowdsourcing as a means to perform the evaluation part thereof (Watts *et al.* 2010; Zen and Gales 2011; Latorre *et al.* 2011; Black *et al.* 2012; Parlikar and Black 2012).

As one example, Hashimoto *et al.* (2011) used MTurk to perform several types of evaluations in the context of speech-to-speech translation (S2ST): intelligibility of TTS, naturalness of TTS, as well as adequacy and fluency of the output of the whole S2ST system and just the machine translation part. Intelligibility was measured by WER while naturalness, adequacy and fluency used MOS. One hundred and fifty people participated in the evaluations, a number well above traditional laboratory experiments. However, crowdsourcing is not the main focus of the paper and no further details of this aspect are given.

So while there is TTS research *using* crowdsourcing, as shown in Chapter 2 there are only a few published studies *about* using crowdsourcing for the assessment of speech technology and even fewer specifically about the assessment of TTS.

The Blizzard Challenges (Black and Tokuda 2005 and newer) evaluate corpus-based TTS systems on common datasets. From the start, they have run evaluations online, and crowdsourced parts of their listeners through an open call for unpaid volunteers and members of participating teams (alongside paid undergraduates in the laboratory). Over the years, slightly different sets of test types have been run, including intelligibility (MRT, SUS), naturalness (MOS, MDS), speaker similarity (DMOS), and appropriateness in dialog context (MOS) (King and Karaiskos 2009). While not mentioned in King and Karaiskos (2010), the organizers of the 2010 Blizzard Challenge repeated the evaluation of one of the main tasks (EH1) on MTurk, and concluded that this is a usable approach (S. King, personal communication).

Wolters *et al.* (2010) conducted tests to measure the intelligibility of four TTS systems using SUS. They repeated the same experiments in two settings: a controlled environment in the laboratory involving 20 university students, and crowdsourced via MTurk involving 159 workers from a variety of age groups and backgrounds. While absolute WERs were worse in the crowdsourced experiments, relative differences between systems were preserved and the much larger number of listeners in the crowdsourced experiments resulted in more statistically significant differences. They conclude that crowdsourcing is viable for TTS intelligibility tests.

Although not for TTS, Chen *et al.* (2009) describe a crowdsourced listening test framework for evaluating algorithms that process multimedia files, such as audio or video codecs. For evaluating n algorithms, the framework runs a task consisting of $\binom{n}{2}$ paired comparisons (*AB* tests). In each comparison, users see/hear the output of one algorithm or the other depending on whether they press or release the space bar. They then submit a forced choice preference

between the two states (pressed/released). The assignment between algorithms and states is randomized. Tests were run in three settings: the laboratory, MTurk, and "an Internet community." In most but not all cases, the laboratory participants yielded higher quality; however, all three groups were reasonably consistent. In terms of cost and participant diversity, the crowdsourced approaches had a clear advantage. Note that the "spacebar switching" presentation method described in Chen *et al.* (2009) is not suitable for TTS listening tests (except potentially for vocoding research) as the outputs of two different (versions of) TTS systems will typically have different timing structure and cannot simply be switched mid-sample.

Ribeiro *et al.* (2011) present an approach to crowdsource MOS naturalness tests for TTS using MTurk. To validate their approach they run MOS tests on part of the 2009 Blizzard Challenge (King and Karaiskos 2009) samples, and show that (1) two subsequent runs produce highly correlated results, that is, good repeatability, and (2) scores from these runs also correlate highly with those of the Blizzard-paid laboratory students, that is, are reliable.

Buchholz and Latorre (2011) ran five TTS *AB* preference tests laboratory-internally as well as on CrowdFlower (http://crowdflower.com/; accessed 17 October 2012) using MTurk workers and compared results. They found that relative preferences were very similar between the two settings and that results in terms of whether or not observed differences are statistically significant were identical. They also observed that the crowdworkers more often expressed a preference (rather than choosing the answer "No preference").

Eyben *et al.* (2012) used crowdsourcing to evaluate the expressiveness of speech, both human and synthetic. They applied unsupervised clustering to the sentences of an audiobook read in often expressive styles and then evaluated that clustering by running *ABX* expressiveness tests on CrowdFlower/MTurk. Listener were asked which of two sound files (*A* or *B*) was more similar in emotion or speaking style to the reference (*X*). The reference was either from the same cluster as *A* or *B*. The authors then used the data from the best performing clustering to build expressive TTS voices and again evaluated the expressiveness using crowdsourced *ABX* tests. While the crowdsourced tests were not formally verified by comparison to noncrowdsourced ones, both the cluster as well as the TTS evaluation showed a clear "winner," indicating that the tests worked.

Nearly all the research mentioned so far was conducted on English TTS voices (American or British). The only exception are the 2008–2010 Blizzard Challenges, in which one of the voices was Mandarin Chinese (Karaiskos *et al.* 2008; King and Karaiskos 2009, 2010).

Other Crowdsourcing for TTS

Apart from in the *assessment* of TTS, crowdsourcing has also been used to prepare resources for the development of the text normalization part of TTS systems (in which "nonstandard tokens" such as abbreviations, numbers, and symbols are expanded into the corresponding words). For example, Pennell and Liu (2010) and Liu *et al.* (2011) asked MTurk workers to list the standard English form of abbreviations found in a Twitter corpus.

7.3.2 Problem and Solutions: Audio on the Web

For all speech applications, one wants to make sure that the audio works at least for the majority of users (operating systems/browsers/audio players). However, as also described in Chapters 4

and 6, delivering audio over the web is not without problems. For TTS assessment in particular, there are additional complications. First, it is imperative that the audio is played uninterrupted; that is, the whole file is downloaded before playing can start. Otherwise, listeners might mistake a temporary lag for a synthesis fault and score the sample much worse than it ought to be. Second, the output of a TTS system is uncompressed audio, such as the PCM WAV format. Therefore, for assessing the quality of a TTS system, one wants to assess the uncompressed audio, rather than a lossily compressed version of it, such as MP3. (Note that as synthesized speech differs in various ways from human speech, and those differences change with changes in the synthesis method, it would be difficult to prove that the choice of WAV vs. MP3 would never influence listening test outcomes for future research versions of a synthesis system.) Third, for listening tests that involve more than one audio sample, one wants to ensure that these are played sequentially, in the intended order. Fourth, one ideally wants to ensure that workers can only submit answers after having played all audio files.

The last two requirements mean that a programmatic way, such as JavaScript, is needed to control the playback of the audio, and a callback from the audio player confirming when a file has played successfully. This, in turn, means that it is not enough to leave audio playback to whatever player happens to be installed in a worker's browser. A typical solution for such a problem involves Flash. However, while Flash supports the MP3 audio format commonly used on the web, it does not natively support WAV. The fourth requirement additionally means that one needs to somehow influence the mechanism by which workers submit their answers to the crowdsourcing platform. However, this part is typically beyond the control of the requester.

The problems described above meant that the first interface (optional playback) described in Buchholz and Latorre (2011) did not fulfill all requirements. Based on QuickTime, which natively supports WAV, it was not completely crossbrowser, and while it did use JavaScript to enable the button for a second audio file only once the first one had played, it could not actually enforce playing of all audio files before a worker could submit their answer. Figure 7.1 shows this interface for a preference test.

The second interface (mandatory playback), developed about a year later, was based on the only Flash-based WAV player (Fedorov 2012) that an extensive web search found. It had better crossbrowser support and through custom-made code also ensured not only that both audio files were played sequentially but also that the answer radio buttons only showed up once the last file had played. As this was marked as a required question, workers could not submit until they had answered it. This is shown in Figure 7.2.

Both interfaces have an additional checkbox that workers can tick if something is wrong with the audio. This was useful as it uncovered the fact that an initial version of the QuickTime interface using a complicated combination of `<object classid="...">` and `<embed>` tags caused problems in about a third of the cases.

Even with the latest interface, there are still a number of technical issues. There is anecdotal evidence that having another Flash player open in another tab of the same Firefox browser can interfere with playing the audio of the listening test. There seem to be issues related to some sampling rates, and the worker forums also recently had reports of audio files taking much longer to load. This might be related to load on the server and needs to be investigated. Finally in rare cases workers still seem to be able to submit answers faster than should be possible when playing all audio. Due to the lack of direct contact between requesters and workers in crowdsourcing, especially when going through an intermediary such as CrowdFlower, knowing how much technical problems affect workers is very difficult. One solution would be

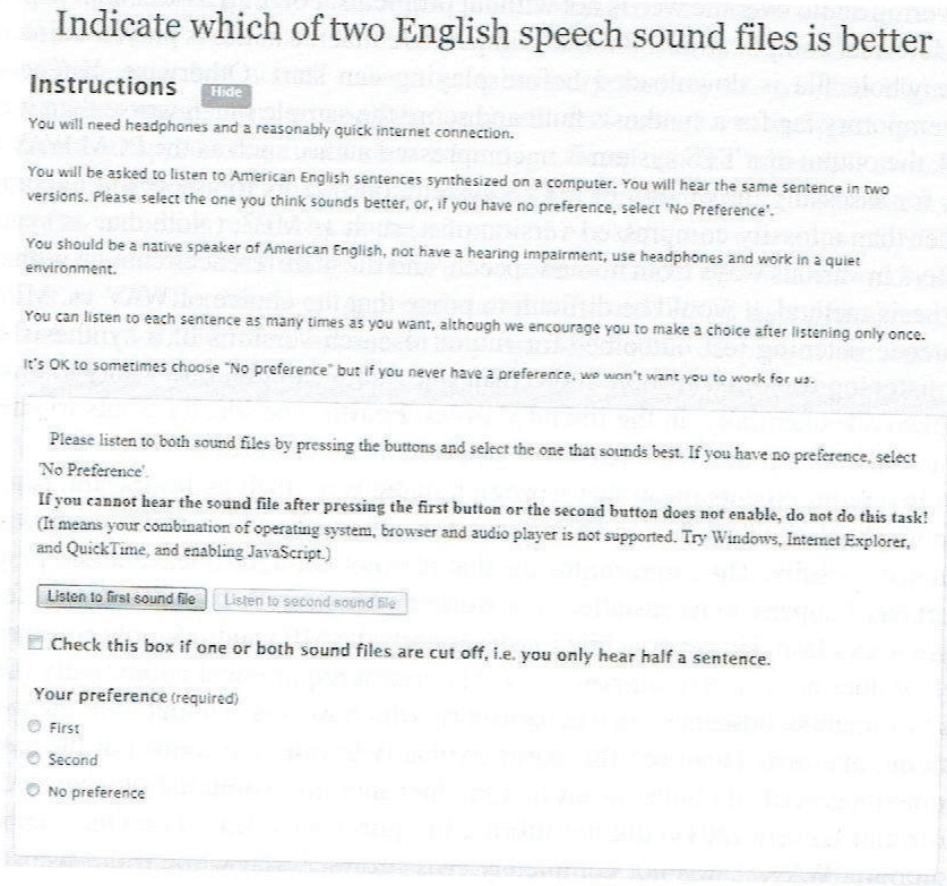

Figure 7.1 The first "optional playback" interface for preference tests.

to include a feedback/comment box into every single HIT (as CrowdFlower does not support a general "exit" questionnaire).

Note that even with the latest development of audio on the web, the HTML5 `<audio>` tag, the problem is not solved. Up to present, different browsers support different subsets of audio formats, and in particular, Internet Explorer does not support WAV (see the overview at http://www.w3schools.com/html5/html5_audio.asp; accessed 17 October 2012).

7.3.3 *Problem and Solution: Test of Significance*

After running a subjective experiment, a test of significance is required to assess how likely it is for the results to be due to chance. This measurement involves comparing the probabilities obtained in the test with the probability of a purely random experiment that is called the null hypothesis. Experiments run in the laboratory are designed in such a way that a standard null hypothesis can be used. In most cases, these standard tests assume that the test is complete: that is, all the desired samples were evaluated. However, in crowdsourcing this cannot be guaranteed. Some files might never be evaluated, some subjects might experience problems, and yet other subjects might later be excluded by quality control measures. To cope with this in the significance test, some adjustments are required.

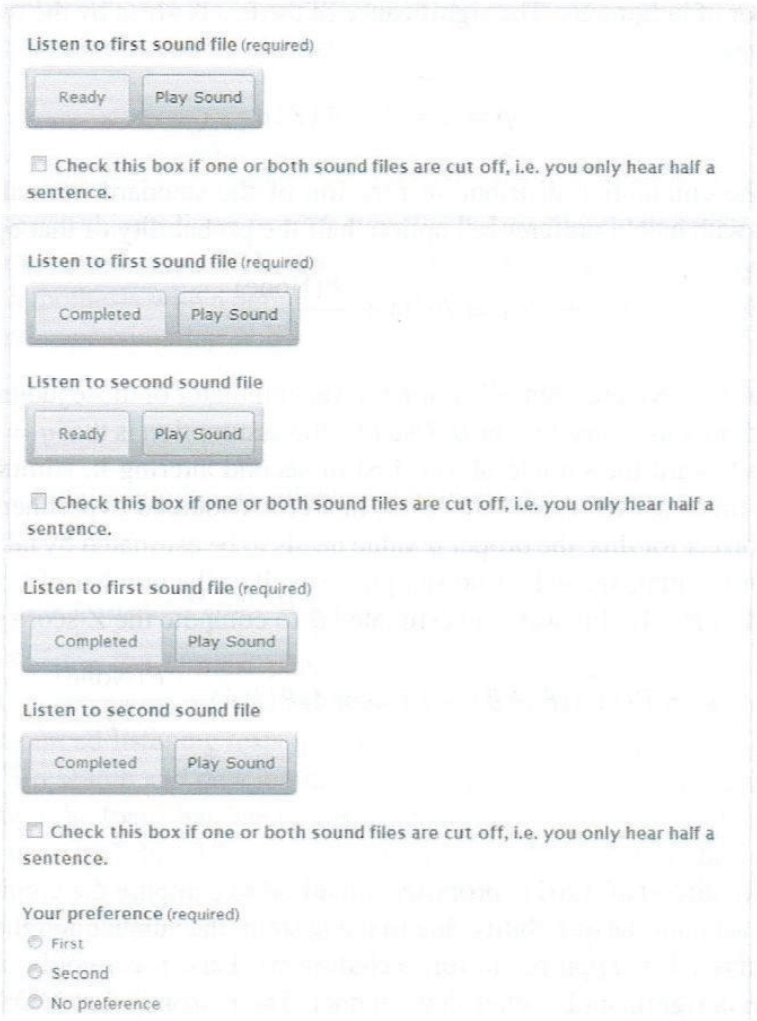

Figure 7.2 The second "mandatory playback" interface for preference tests. The three screenshots show successive stages of the interface as a worker plays one audio after the other.

Comparison Tests

In preference tests, the standard null hypothesis is that the samples are obtained from a Bernoulli distribution with $q = 0.5$. In other words, the null hypothesis is that both systems are equally preferred. In this way, the z-score is computed as the number of standard errors S between the observed system preference X_A and the expected one q:

$$Z = \frac{X_A - q}{S}, \tag{7.1}$$

where

$$S = \frac{\sigma_{\text{bernoulli}}}{\sqrt{N}} = \frac{\sqrt{q * (1 - q)}}{\sqrt{N}} \tag{7.2}$$

with N the number of judgments. The significance of the test is given by the two-tailed p-score that is computed as

$$p = 2 * (1 - \Phi(Z)), \tag{7.3}$$

where $\Phi(Z)$ is the cumulative distribution function of the standard normal distribution. In comparison tests with a "No preference" option, half the probability of that option is added:

$$X_A = P(A) + \frac{P(\text{None})}{2}. \tag{7.4}$$

This is justified, as the "No preference" option is an accumulator of those judgments that would otherwise be randomly assigned to A or B. Usually, the assumption is that $q = 0.5$ and to avoid any possible bias toward the sample played first or second altering it, stimuli are played the same number of times as AB as BA. However, in a crowdsourced experiment this cannot be guaranteed. To correct for this, the proper q value needs to be estimated by taking into account the preference for the first, second, or no sample, as well as the number of times stimuli were played as AB and as BA. In this way, the estimated $\hat{q}$ to compute the Z-score should be

$$\hat{q} = P(\text{first})P(AB) + P(\text{second})P(BA) + \frac{P(\text{None})}{2}. \tag{7.5}$$

MOS Test

For MOS tests, Ribeiro *et al.* (2011) proposed a method to compute the confidence score that tries to take into account the variability due to the system, the subject and the subjective test. Although this approach is appropriate for excluding workers, it cannot be used to compare whether a system is significantly better than another. The reason is that MOS is a Likert-type scale (Likert 1932), which means that although the MOS scale is ordered, the distance between two consecutive levels is not guaranteed to be equal. As a result, it is not appropriate to compare systems based on their mean MOS scores. This problem was discussed in Clark *et al.* (2007), in which the use of a Bonferroni-corrected Wilcoxon signed rank test was proposed. The problem with this type of test is that it assumes a set of matched results for the compared systems, but this condition is not always true for crowdsourced experiments. A possible way to solve this is to consider the MOS test as a kind of asynchronous comparison test and test the statistical significance of $P(A > B)$ versus the null hypothesis that the scores for A and B come from the same distribution. Assuming statistical independency $P(A = i, B = j) = P(A = i)P(B = j)$, therefore, for a 5-points MOS scale,

$$P(A > B) = \sum_{i=2}^{5} \sum_{j=1}^{i-1} P(A = i)P(B = j) \tag{7.6}$$

and

$$P(A = B) = \sum_{i=1}^{5} P(A = i)P(B = i). \tag{7.7}$$

Equating these probabilities to $P(A)$ and $P(\text{None})$ of a standard comparison test, respectively, the preference for system A can be estimated as

$$\bar{X}_A = P(A > B) + \frac{P(A = B)}{2}. \tag{7.8}$$

Since in a MOS tests subjects listen to each stimulus independently and the presentation order is randomized, it is almost impossible to estimate the corrected $\hat{q}$. Therefore, it is reasonable to assume a null hypothesis with a default $\hat{q} = 0.5$. The standard error is then computed using Equation (7.2) and approximating N by

$$\bar{N} = \sqrt{N_A B_B}, \tag{7.9}$$

where N_A and N_B are the number of collected judgments for systems A and B, respectively.

7.3.4 *What Assessment Types Worked*

As was seen in Section 7.3.1, most papers about crowdsourced assessment of TTS report about one or two experiments, performed at one point in time, for English only. At Toshiba Research, crowdsourced listening tests have been used routinely since the middle of 2010 to evaluate new TTS research and development. A variety of test types was run, for a variety of languages, although the focus has been on English. Table 7.2 gives an overview.

The listening tests in Table 7.2 were all run on CrowdFlower (another platform will be discussed in Section 7.3.6). CrowdFlower offers several "labor channels"; that is, they can post tasks on different platforms. Unless otherwise noted, the listening tests discussed in the remainder of this chapter used the MTurk labor channel; that is, CrowdFlower posted the tasks on MTurk where they were completed by MTurk workers. Officially, only people or institutions located in the United States can directly post work on MTurk (MTurk 2012a). This is the main motivation for using CrowdFlower, as they do not impose such a restriction. They also offer easy support for using a gold standard (see Snow *et al.* 2008) as quality control *during* the task. Workers are prevented from continuing work on the task as soon as they fail too many gold questions. However, unless otherwise stated, this functionality was not used, due to the subjective nature of listening tests. Finally, while MTurk offers the possibility to

Table 7.2 Numbers of listening tests (run via CrowdFlower), by test type and language.

Test\language	enUS	enUK	deDE	esUS	esES	frFR	frCA	All
AB	229	9	32	4	4	5	0	283
ABX	54	0	0	0	0	0	0	54
MOS	32	2	4	5	0	2	1	46
DMOS	12	0	0	0	0	0	0	12
total	327	11	40	9	4	7	1	400

enUS, American English; enUK, British English; deDE, German (Germany); esUS, American Spanish; esES, European Spanish; frFR, European French; frCA, Canadian French.

restrict access to a task to workers from specified countries, it does so on the basis of the country that the *worker* filled in when signing up. CrowdFlower, on the other hand, uses IP address geolocation to actively prevent access to tasks for workers who are not located in the countries specified by the requester. This is a welcome extra advantage.

However, there are also disadvantages: CrowdFlower charges a percentage on top of MTurk's price; did not initially offer the possibility to block specific workers from all future tasks; and does not offer functionality for rejecting workers, paying bonuses or external HITs (where the task is hosted on the requester's server rather than Amazon's). The latter is of particular relevance to listening tests. Wolters *et al.* (2010) ran each complete intelligibility test (involving several questionnaires and 50 SUS samples) as a single human intelligence task (HIT) on MTurk by using the external HIT interface. This gives complete control over the experimental setup and allows advanced test designs. For example, the Blizzard Challenges use a Latin square design for their listening tests (Bennett and Black 2006). This is not possible on CrowdFlower. All listening tests discussed here were, therefore, run in such a way that each rating (MOS, DMOS) or paired comparison (*AB*, *ABX*) was its own HIT. This setup probably speeds up completion rates as it is less risky for workers who do not yet know this type of task to sample a few (as they can just stop at any point and be paid proportionally). Wolters *et al.* (2010) reported having to re-release the task each day for at least 8 days, while the English listening tests discussed in this section mostly finish within a day.

The remainder of this section discusses in more detail how the individual listening test types were implemented.

MOS Tests: Naturalness, Quality, Diagnostic

A number of types of MOS tests have been run: MOS *naturalness*, MOS *quality*, and *diagnostic* MOS. The first two types are rated on a 5-point scale, as described in Section 7.2. MOS naturalness is rated on a continuum of "completely natural" – "completely unnatural," whereas the continuum for MOS quality is from "very good" to "very bad." These have been used to evaluate systems for research purposes as well as part of quality assurance in product development, whereby our own system is compared against those of the competitors. The third type, diagnostic MOS, is a modified form of MOS, in which the objective is to examine the extent of any problems in a large number of sentences, typically 2000 or more. The ratings are on a 3-point scale, and the points are labeled "No problem," "Minor problem," and "Major problem."

In ordinary MOS tests, the overall performance of the systems is our main interest, and as such, scores for individual sentences are irrelevant. In diagnostic MOS, however, scores for individual sentences are respected. This is used as a filter for a more detailed diagnostic evaluation, which is described in Section 7.3.5. Note that in order to draw the listener's attention to any linguistic or phonetic errors as well as synthetic artifacts, the text of the sentence is displayed in the interface for diagnostic MOS.

In all diagnostic MOS tests run on CrowdFlower, gold-standard items were included. They consisted of "bad" gold, in which problems were introduced artificially, as well as (in later tests) "good" gold, which were (1) unanimously judged to be problem free in a gold-selection diagnostic MOS test that preceded the main test and then (2) confirmed as being incontestably "good" by an in-house native speaker of the language. A small number of MOS quality and MOS naturalness tests were run with human speech samples as gold standard items to anchor the upper limit of the scale, but the majority were run without gold.

CrowdFlower/MTurk ensures that workers are served HITs in random order and that they are never served the same HIT more than once. However, as already pointed out in Section 7.3.4, it has not been possible to implement tests with a Latin square design using CrowdFlower, so a listener may hear the same sentence (by different systems) more than once.

Listeners need to listen to a few samples before they are able to make a call on the range of naturalness, quality, or extent of errors present in the evaluation samples. Therefore, the first three responses from each worker are normally excluded from the analysis of MOS tests.

Since June 2010, 64 MOS tests have been run on CrowdFlower/MTurk and Clickworker (http://www.clickworker.com/; accessed 17 October 2012), out of which nine are diagnostic MOS. The tests cover seven languages, completed or aborted with varying degrees of success. A worker is paid US$0.03 per HIT. This amount is based on the US federal minimum wage and the average time it is expected to take for a worker to complete a HIT.

AB Tests: Preference

As explained above, each HIT of an *AB* preference test consists of just one paired comparison. Workers see instructions and two buttons that they need to click to play the two audio samples, and then indicate their preference for the first sample, the second sample, or none (cf. Figure 7.1). To avoid presentation order bias (Wherry 1938), separate HITs are posted for each order (*AB* as well as *BA*) of each sample pair.

For each preference test, the test outcome is computed in terms of percentages of preference expressed for each system, *A* and *B*, and "No preference." To check whether any observed difference in preference for the systems is statistically significant, the two-tailed *t*-test described in Section 7.3.3 is applied to the results after splitting the "No preference" votes equally over the two systems, simulating a forced choice situation in which people who really have no preference can be expected to vote for *A* as often as for *B*.

The above method to crowdsource preference tests has been used within our TTS research group extensively. Since June 2010, 284 preference tests have been run on CrowdFlower. Research evaluated encompassed a wide range of topics, including prosody, acoustic modeling, and vocoding. Typically, all but the largest test groups finish overnight, which greatly increases researcher efficiency.

By default, a test consists of a minimum of 50 sample pairs played in both orders, and each pair is evaluated by at least 5 workers, yielding a total of 500 judgments. Usually each worker is allowed to provide a maximum of 40 judgments (functionality provided by CrowdFlower). Therefore, the minimum number of workers per test is 12.5. A worker is paid US$0.05 per comparison and the average sound sample is 8.9 seconds long. US workers so far came from all the 50 states and, therefore, should be more representative than any group one could hope to assemble in a traditional laboratory experiment.

DMOS and ABX Tests: Similarity

DMOS and *ABX* tests were used for testing the similarity to a given sample in terms of speaker identity by Zen *et al.* (2012) and Wan *et al.* (2012), and in terms of expression/speaking style by Eyben *et al.* (2012), Latorre *et al.* (2012), and Chen *et al.* (2012). As these tests involve listening to more audio samples than for an *AB* preference test, a worker is paid US$0.05 and US$0.07 per comparison, respectively. We have not carried out any specific analyses for these test types other than the standard analysis used for standard MOS or *AB* tests. In general, subjects in these types of tests find it difficult to abstract the factor for which they

are asked from the general speech quality of the samples. People who are used to listening to TTS systems may have learned to discriminate between the different factors, but this is not necessarily the case for naive listeners. Even in controlled experiments, when comparing speaker similarity in a DMOS test with a human speech sample as the reference, it is not unusual for listeners to give higher scores to the samples that "sound" better, even though the speaker identity might actually be very different. To avoid this, the question for these types of tests has to be formulated very carefully so as to be as unambiguous as possible. For crowdsourced experiments this is even more important, as the average naivety of the workers tends to be higher.

7.3.5 What Did Not Work

While the listening test types described above could be crowdsourced successfully, there were some types that did not work well. Here, we will describe two experiments that we failed to crowdsource successfully.

Diagnostic Evaluation

One such type is a diagnostic evaluation, which is designed to pin-point problems of quality in a set of sentences. The desired result is not a MOS or a preference, but a set of indications concerning problems of certain kinds such as synthesis artifacts, text normalization errors, mispronunciations, pause errors, and intonation errors. While also meant to identify segmental errors, the scope of a diagnostic evaluation is much broader than the segmental tests discussed in Section 7.2. It can serve as a tool for quality assurance purposes, highlight the areas that need improvement, and provide per-word diagnostics of sentences agreed upon by multiple listeners. By combining the diagnostic results with MOS scores, it may also be possible to obtain an indication of which error types have a greater impact on the overall quality than others.

Previously, attempts were made to collect such information from free-form comments. For example, in an experiment conducted with naive listeners in a controlled environment, Krstulović *et al.* (2008) encouraged listeners to give free-form comments for each sentence, in order to augment the information collected by means of MOS scores. The comments were then manually reduced to a set of semantic tags to summarize the problem in a standardized manner. However, this involved a potentially costly phase of manual lexical and semantic analysis, and would be impractical with a large number of utterances and listeners. While one option would be to crowdsource this tagging phase, we opted for an approach in which the listeners' judgments are obtained in a more direct manner, and in finer detail.

A pilot experiment for a diagnostic evaluation was run on CrowdFlower/MTurk with 20 evaluation sentences and 10 gold-standard sentences. The gold-standard sentences consisted of "problematic" sentences, in which problems were artificially introduced. Problems included mispronunciations (segmental and suprasegmental), pause insertion, cutoff, flat prosody, and overlaying artifacts such as echo, noise, and beeps. These were intended to serve as training examples as well as to exclude spammers. CrowdFlower allows, upon request, for a certain number of Gold samples to be presented at the beginning of the evaluation for each worker, so that any workers who fail these are barred from proceeding further.

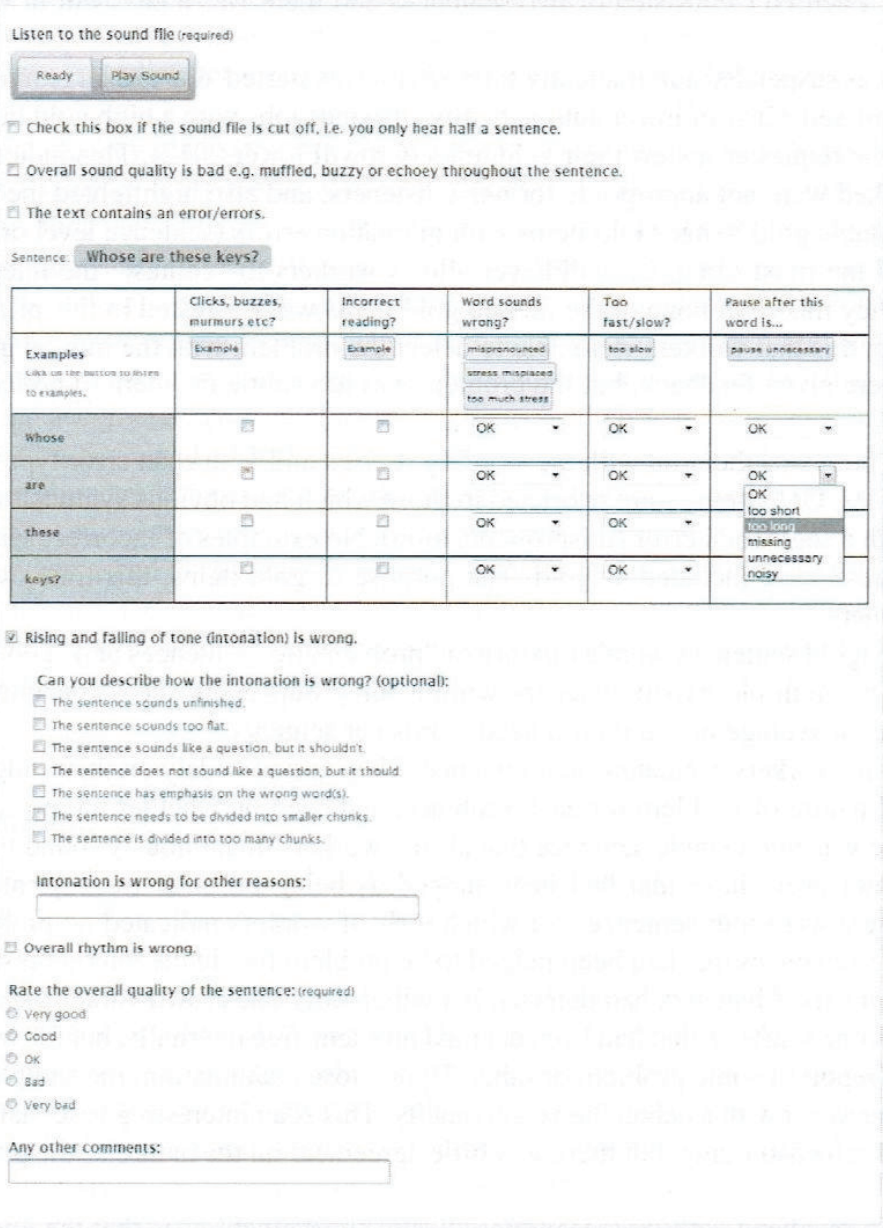

Figure 7.3 A screenshot of the diagnostic evaluation interface.

Figure 7.3 shows an interface similar to the one used. Listeners were presented with the unnormalized text of the sentence as they listened to the corresponding audio file. They were asked to indicate any problems at the word level, by ticking the checkbox or selecting an option from the dropdown menu in the intersection cell for the word and the error type. Word-level errors included artifacts, text normalization, mispronunciation, lexical stress, sentence stress, rhythm, and pause. They were also asked some questions about sentence-level errors such as overall intonation contour, chunking, and overall sound quality. Each error type was described in a nontechnical manner and an explanation was given in the instructions. Finally, listeners were also asked to indicate the overall quality of the sentence on a scale of 1 (very bad) to 5