Intelligibility of Machine Translation Output in Speech Synthesis

Laura Mayfield Tomokiyo, Kay Peterson, Alan W. Black, Kevin A. Lenzo

Cepstral LLC
Pittsburgh, PA USA
laura@cepstral.com

Abstract

One use of text-to-speech synthesis (TTS) is as a component of speech-to-speech translation systems. The output of automatic machine translation (MT) can vary widely in quality, however. A synthetic voice that is extremely intelligible on naturally-occurring text may be far less intelligible when asked to render text that is automatically generated. In this paper, we compare the quality of synthesis of naturally-occurring text and its MT counterpart. We find that intelligibility of TTS on MT output is significantly lower than on either naturally-occurring text or semantically unpredictable sentences, and explore the reasons why.

Index Terms: Speech synthesis, Speech-to-speech translation, TTS evaluation, TTS intelligibility

1. Introduction

Although text-to-speech synthesis (TTS) is often used as the end component of speech-to-speech translation systems, the performance of synthetic voices on the output of the machine translation (MT) component is not well understood. TTS voices are generally designed to perform well on a specific type of input. Although the input may be limited in domain, sentence-level text that is sent to a TTS engine is almost always expected to be well-formed.

The process of machine translation can be thought of as a noisy channel that degrades the quality of the text that is sent to synthesis. It generates text that can be grammatically incorrect, fragmented, or telegraphic. Standard measures of machine translation quality assess whether the meaning is conveyed, how many words are incorrect, or how many words it would take to change the sentence into a meaningful one. We are not aware of any recognized metrics that take into account appropriateness for TTS.

What we strive to quantify in this paper is the difference in intelligibility between synthesis of naturally-occurring text and synthesis of MT system output, for a unit selection synthesis engine. We simulate the MT channel using a major online translation engine. We then evaluate TTS using transcription accuracy, smoothness, and subjective quality metrics on the source and translated text, assessing performance and placing TTS of MT output on a continuum that includes semantically unpredictable sentences and random word sequences.

2. Background

The study described in this paper was motivated by our experiences with an English-Arabic speech translation system, developed as part of the DARPA TRANSTAC program. It was observed that listener judgements of the Arabic synthesis were substantially more negative for sentences that were generated as part of the end-to-end system than the clean test sentences that we had worked with in development.

While speech synthesis is viewed as an integral part of speech-to-speech translation systems, end-to-end evaluation of the system is actually often performed at the text level, using scoring metrics such as BLEU [1]. Evaluation of the TTS module is often done in isolation or not mentioned at all (e.g. [2]). The ultimate objective of a speech-to-speech translation system, however, is to provide spoken output in contexts where display or comprehension of text output is not possible, and a clear understanding of performance of TTS under different system conditions is necessary for the evaluation of the system.

Evaluation of TTS even in isolation is difficult, and necessarily subjective. The Blizzard challenge [3] is one of the first large-scale efforts to evaluate and compare quality of synthesis across systems in a controlled setting, similar to the speech-to-speech system evaluations for projects like Verbmobil, Nespole!, and TC-STAR.

It has been observed that semantic content can affect perceived quality of speech under noisy signal conditions (e.g. [4]). Performance on Semantically Unpredictable Sentences (SUS) is also recognized as a valuable factor in assessing TTS quality [5]. The "SUS channel" does not quite approximate the "MT channel," however, in the nature of the noise introduced. Depending on the type of MT, we may see sentence fragments, word-for-word translations, or untranslated words in the input. We may see, instead of the appropriate target-language term, a term that has a similar distribution but that completely changes the syntax or semantics of the sentence. Or, realistically, we may see a combination of these and other types of noise in the channel.

The output generated by MT systems is certainly not random, however. Even a poorly translated sentence is often understandable when the reader has the opportunity to dissect it visually. The question that we would like to answer is how much TTS intelligibility degrades when the input is not naturally-occurring, well-formed text, but rather the output of an MT system.

3. Experimental Framework

We considered three aspects of intelligibility in our experimental design: transcription accuracy, smoothness of synthesis, and overall subjective impression. Ratings could easily vary along any one of these dimensions while the other two remain the same. For example, the listener might understand each of the words and also hear each word as being smoothly synthesized, but just feel that the quality of synthesis is poor, perhaps because the prosody is unnatural. Alternatively, the synthesis could be completely smooth and sound natural, but the listener thinks they have heard something else, and transcribes the sentence incorrectly. This can be the case when the word sequence itself does not fit the listener's internal language model, as with semantically unpredictable sentences or "noisy" MT output.

This section provides a general description of the test data and

methodology. Detailed results are given in Section 4.

3.1. Test Environment

3.1.1. Synthesis

Cepstral's Swift™ TTS engine was used for all experiments. Swift is a unit selection synthesizer. The voice used was Cepstral's David voice. David is a US English voice, and is consistently judged to be among Cepstral's highest-quality voices for typical use.

3.1.2. Web evaluation

Test participants (subjects) took the tests remotely, on their own computer equipment. We did not instruct subjects to use headphones or not, or impose any other hardware or surrounding requirements, as these would have been difficult to control.

When clicking on a link, subjects were taken to the initial page of a test in a web browser, where they had to enter their unique username (which was provided to them). After entering their username, they were shown the first page of the respective test.

For the transcription tests, subjects saw a screen of links to audio files, with a space to put in their transcription of the audio next to the audio link. Clicking the audio link allowed them to listen to the sentence and then transcribe it. They were allowed to listen to the sentences as many times as they wished.

For the smoothness and subjective quality tests, subjects were presented with a page with an audio link for one sentence and the words from that sentence displayed next to the link. Subjects were asked to first listen once without looking at the text and enter their overall subjective quality rating in a box to the left of the text. They then were to listen to the sentence again looking at the text. Every word could be marked individually for non-smooth synthesis. Subjects listened to the audio and then marked the words considered non-smooth.

3.2. Participants

3.2.1. Recruitment

Participants in the experiment were recruited via an academic experiment website and word of mouth and were compensated in cash or with text-to-speech synthesis software.

3.2.2. Demographics

A total of 17 subjects completed the study. All of them were native speakers of US English from a variety of US regions, with a concentration in the North-East regions. Subject age ranged from 18 to 40, with the majority of subjects being in the 18-25 age bracket. Most of them were college students. Gender was not recorded. None of the subjects had any prior experience working with text-to-speech synthesis.

3.3. Input text

Synthesis on four types of text was examined: naturally-occurring text, MT output, random-word sentences, and semantically unpredictable sentences. All sentences were between 5 and 12 words and the average length per set was 8 or 9 words per sentence.

3.3.1. Naturally-occurring text

The naturally-occurring text was taken from the English segment of the European Parliament (Europarl) [6] transcriptions. This data closely resembles some of the data on which the TTS voice was trained, and using well-matched data minimizes confounding factors in synthesis quality. Examples of naturally-occurring text follow.

- We have spent two years extensively discussing the Community fisheries market.
- This distinguished colleague endeavoured to improve the Commission's proposals.

3.3.2. MT output

To generate text that would simulate output from an MT system, we used a major online translation program to translate the naturally-occurring text out of English and back again. We attempted to control the translation quality using recognized measures such as BLEU score, but BLEU is not appropriate for a small number of samples (we used 10-20), and measuring BLEU score in isolation did not give us a good measure of badness. Ultimately, we felt that a subjective assessment of the translation output gave us the greatest ability to match translation quality to that which we had seen in speech-to-speech translation systems. Examples of MT output text corresponding to the above examples for naturally-occurring text follow.

- We largely spent two years the examining Community market of fishing.
- This remarkable colleague disturbed to improve the requests to him of the Commission.

3.3.3. Random-word sentences

Although the MT output data described above is somewhat noisy, it generally has some syntactic coherence. Semantically, we see both unpredictable and predictable words. The question that one must ask is whether the semi-sound sentences in the MT data are closer in intelligibility to well-formed sentences or to random sequences of words in a similar domain.

To test this, we took a separate set of sentences from the Europarl data, and generated 8-word sentences by randomly taking words out of all tokens. Specifically, we listed all the tokens in a 232-word text, randomized that word list, and formed sentence 1 out of words 1-8, sentence 2 out of words 9-16, and so forth. The vocabulary distribution was similar to the naturally-occurring and MT sentences, but there was no syntactic or semantic structure other than what appeared coincidentally. Examples of random-word sentences follow.

- At mandate report concentrated the our lowered view.
- The Indonesian there I very matter council the.

3.3.4. Semantically Unpredictable Sentences (SUS)

Semantically Unpredictable Sentence (SUS) tests are a common tool for isolating synthesis quality from word context. To create a SUS test, a number of sentence patterns are defined with slots that can be filled, and these slots are instantiated with random selections from large lists of words of the appropriate part of speech. Examples of semantically unpredictable sentences follow.

- The copier bleeds heavily on a fill.
- Another purse inside a basement relaxes that same soda water.

3.4. Test Methodology

Three types of tests were used for evaluation: transcription tests, smoothness tests, and subjective quality tests.

3.5. Transcription tests

In the transcription test, listeners hear a synthesized utterance and are asked to transcribe exactly what they hear. This test evaluates how well the listener can understand the synthesized speech. Listeners must transcribe words they understand even if they think the synthesis is poor, so the transcription test provides a fairly objective measure of the bottom line – whether or not the correct meaning is conveyed.

Participants in this study completed transcription tests for each of the data sets - naturally occurring text, MT output, random-word sentences, and SUS. Because we wished to compare transcription results for the same source text before and after the MT step, and as the transcription task gets easier the more times listeners hear the audio, we divided the participants into two groups (group 1 and group 2). Two texts were prepared (text A and text B). Texts A and B were then passed through the MT channel to generate text AT and text BT. Listener group 1 transcribed text A and text BT, while listener group 2 transcribed text B and text AT. Results from clean sets A and B were then averaged to obtain a score for transcription of clean text, and results from MT sets AT and BT were averaged to obtain a score for transcription of MT output. Both groups transcribed the random-word set CR and the semantically unpredictable set SUS. All test sets contained 19 utterances, except for the SUS set which contained 10 utterances.

In background analysis, the similarity in performance of groups 1 and 2 for the sets they had in common (CR and SUS) indicates that agreement between the groups is good, and any difference in accuracy on the two data sets A and B is because of a difference in the data. It does appear that set B was more difficult than set A; transcription accuracy is substantially worse both before and after the MT step.

3.6. Smoothness tests

In the smoothness test, listeners are shown a sentence and are asked to mark each word that does not sound smooth. They are permitted to listen to the audio as many times as they wish.

The smoothness test gives us a measure of how smooth the synthesis is, but more importantly, in conjunction with the transcription test it can give us a sense of how important synthesis join errors are. If two data sets score similarly on the transcription test but differently on the smoothness test, it tells us that join errors are noticeable but do not affect understanding. Conversely, if two tests score similarly on the smoothness test but differently on the transcription test, we may conclude that problems in understanding are not directly attributable to join errors.

3.7. Subjective quality tests

In the subjective quality test, listeners hear an utterance and give it an overall quality score of 1-5. This rating is sometimes known as a Mean Opinion Score (MOS). There are arguments both for and against allowing the listener to see the text being synthesized when they give their rating; our listeners were asked not to look at the text. Generally, a score of 1 is unintelligible and a score of 5 is both highly intelligible and natural-sounding.

4. Results

4.1. Transcription tests

Scoring of transcription tests was done using the NIST sclite [7] scoring tool to calculate a word accuracy rate identical to that used to assess speech recognition output. A few mapping rules were applied to allow some flexibility in truly ambiguous situations (e.g. "St." vs. "street"; US "favor" preferred by listeners vs. UK "favour" found in the reference data). No correction was made to misspelled words in the listener transcriptions. It has been our experience that the impact of occasional spelling errors on the accuracy score is small, and tends to balance out across data sets.

Text Source	Transcription Accuracy (%)
Naturally-occurring text	93.5
MT output	86.7
SUS	77.8
Random sentences	62.2

Table 1: Transcription accuracy

Results of transcription tests are given in Table 1. The loss in transcription accuracy in MT data is 6.8% absolute (7.3% relative). We also see that transcription accuracy is higher for the MT output than for the semantically unpredictable sentences, suggesting that there is still a lot of information in the MT sentences that helps with understanding.

4.2. Smoothness tests

Text Source	Transcription Accuracy (%)
Naturally-occurring text	91.8
MT output	90.1

Table 2: Smoothness

In this study, smoothness tests were only conducted for the natural-MT comparison. Results are shown in Table 2. We see that although the smoothness of the MT output is somewhat lower than for the naturally-occurring text, the difference is not as striking as it is for the transcription tests.

4.3. Subjective quality tests

Clearly, there is much room for interpretation on the part of the listener. The subjective quality score does, however, give us a sense of the listener's perception of quality that we may not get from the two objective tests.

Text Source	MOS score
Naturally-occurring text	4.4
MT output	4.1

Table 3: Subjective quality (1-5 scale)

Results for the subjective quality test are shown in Table 3. Again, only naturally-occurring text and MT output were scored. We see a continunation of the trend that the perception of quality in synthesis of the MT data is lower than for synthesis of naturally-occurring text.

5. Potential Solutions

The results outlined above show that intelligibility scores of TTS on MT output are lower than those of naturally-occurring text. The question we must attempt to address is how this effect can be ameliorated in a speech-to-speech translation system, where synthesis of MT output is the entire objective of the system.

5.1. Tight coupling

Where the system architecture allows it, MT hypotheses can be reordered using a measure of predicted synthesis quality[8]. For example, a sentence that is well predicted by a word or phoneme language model trained on the TTS training data might receive a high score. Sentences with words outside the TTS vocabulary might receive a low score. Alternatively, the system might maintain a list of acceptable word replacements, and substitute a word that is known to be synthesized well for one in the MT hypothesis that is not.

This concept is similar to strategies learners of a language use to choose the words they know how to say well.

5.2. Modification of synthesis

If feedback from the upcoming synthesis step cannot be used in selection of the MT hypothesis, the synthesis engine itself can be more sensitive to the quality of the input and enhance the spoken output in a way that makes it easier to understand. Preliminary experiments have shown that for noisy text, slowing the overall speaking rate, and more importantly, adding and extending phrase breaks within a sentence, can improve intelligibility of synthesis.

We can also find a model for this strategy in human language, with speech directed toward non-native listeners. Speakers that are accustomed to interacting with listeners of limited listening comprehension skills often speak slowly and clearly, and break frequently to allow processing time. Human speakers reading aloud also read noisy text differently from clean text, perhaps both for the listener's sake and because articulation of the unexpected sequences is difficult.

It would also be possible to add a word substitution step at synthesis time, although there may be a greater danger of using an inappropriate word sense when outside the MT loop.

6. Discussion

There are a number of explanations for the observation that MT output is less intelligible in synthesis than naturally-occurring text. MT output can contain word sequences that are very unpredictable, and would be difficult to transcribe without visual cues even from clearly read speech. The language model and auditory recovery strategies that native listeners employ can trick them into thinking they hear something different than what was actually spoken.

One reason that synthesis is less smooth on unusual word sequences is that the synthesis engine is optimized for the word and phone sequences that occur in the TTS training and development data. MT output may break some of the "rules" of human language generation, juxtaposing words that are difficult to pronounce together. Just as human speakers can stumble over unexpected word combinations, a synthetic voice can encounter more join errors and data sparsity problems.

7. Conclusion

In this paper, we have attempted to quantify the loss in TTS quality that occurs when the synthesis is used as a back end to machine translation, as in a speech-to-speech translation system. We have compared synthesis of naturally-occurring text and MT output along three dimensions: transcription accuracy, smoothness, and subjective quality. In all cases, synthesis of MT output was judged to be inferior to synthesis of naturally-occurring text. MT output appears to fall in between clean text and semantically unpredictable sentences in terms of intelligibility in synthesis.

The experiments in this paper also examined the effects of other noisy channels on TTS intelligibility. Semantically unpredictable sentences were less accurately transcribed than MT output, and random sequences of in-domain words were even harder to understand. This suggests that there is a hierarchy of corruptions that can damage TTS quality. If we contrast the SUS data with the MT output, and quantify the quality of the text in terms of something like the minimal edit distance to produce a meaningful sentence, the SUS data may score higher, yet the MT data is still more intelligible. Similarly, while an MT-output sentence like "Where sport, you arrive above to which tops" may not seem to different from a random sentence like "Need appropriation on this for one will taxes," there is evidently something about the MT data that makes it more intelligible in synthesis. Deeper examination of this hierarchy is left for future work.

8. Acknowledgements

This research is sponsored in part by the Defense Advanced Research Projects Agency (DARPA) program "Spoken Language Communication and Translation Systems for Tactical Use (TRANSTAC)", under contract ANBCHC0300280002.

9. References

[1] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, "BLEU: a method for automatic evaluation of machine translation," Tech. Rep. rc22176 (w0109-022), IBM, IBM Research Division, Thomas J. Watson Research Center, 2001.

[2] Alon Lavie, Florian Metze, Roldano Cattoni, and Erica Costantini, "A Multi-Perspective Evaluation of the NESPOLE! Speech-to-Speech Translation System," in *Proc. ACL Workshop on Speech-to-Speech Translation: Algorithms and Systems*, 2002.

[3] Christina Bennett, "Large Scale Evaluation of Corpus-based Synthesizers: Results and Lessons from the Blizzard Challenge 2005," in *Proc. Interspeech*, Lisbon, 2005.

[4] Alexander Raake, "Does the Content of Speech Influence its Perceived Sound Quality," in *Proc. LREC*, 2002.

[5] Christian Benoît, Martine Grice, and Valérie Hazan, "The SUS Test: A Method for the Assessment of Text-to-Speech Synthesis Intelligibility using Semantically Unpredictable Sentences," *Speech Communication*, vol. 18, no. 4, pp. 381–392, 1996.

[6] Philipp Koehn, "Europarl: A Multilingual Corpus for Evaluation of Machine Translation," http://people.csail.mit.edu/koehn/publications/europarl.ps, 2002.

[7] NIST, "Speech Recognition Scoring Toolkit (SCTK)," 2000, http://www.nist.gov/speech/tools/.

[8] Tanja Schultz, Alan W. Black, Stephan Vogel, and Monika Woszczyna, "Flexible Speech Translation Systems," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 2, 2006.