VISUAL EVALUATION OF VOICE TRANSFORMATION BASED ON KNOWLEDGE OF SPEAKER

Arthur R. Toth and Alan W Black

Language Technologies Institute
Carnegie Mellon University, Pittsburgh, PA, USA
atoth@cs.cmu.edu, awb@cs.cmu.edu

ABSTRACT

Voice transformation techniques are maturing. The ability to automatically change a source voice to a target voice, using a model built from a small amount of target speaker data, has brought with it the need to better understand how to evaluate the quality of a transformation model. This paper presents a simple experiment to measure how familiarity with the particular source and target speakers affects perception of the transformation. The results show that listeners' views of the transformation are not affected by familiarity with the speakers. In addition to these results, we also introduce **Transformation Triangle Diagrams**, a graphical mechanism to better display certain relationships that are important in the evaluation of voice transformation.

1. INTRODUCTION

Voice transformation is one of a number of names given to techniques which take speech from a source speaker as an input and attempt to produce speech that sounds like a target speaker. These techniques have been used in a number of applications [1]. One compelling argument for studying voice transformation is that it may reduce the difficulty in creating new voices for speech synthesis. Once a full-sized corpus has been collected for a source speaker, the amount of additional data needed to produce a new voice with a typical voice transformation system is much smaller than what is necessary to produce a new voice based on concatenative synthesis alone.

One natural question to ask about voice transformation techniques is how to measure their quality. Intelligibility, naturalness, and speaker recognizability are factors that are commonly measured in the assessment of voice transformation quality [2]. Furthermore, attempts to measure these factors consist both of "objective" and "subjective" tests [2]. Objective tests provide metrics that can be calculated from the output speech and reference speech directly. Subjective tests involve collecting opinions from people in listening experiments and analyzing the results. The strength of objective tests is that they can be performed quickly and automatically. However, when it comes to measuring the quality of voice transformation, the "gold standard" is human perception, and subjective tests are based on it. When objective tests are employed, they are typically used in conjunction with subjective tests and some attempt to correlate the results of the tests is used to justify the objective tests.

Although subjective listening tests have the great advantage of being based on human perception, they are, at their base, subjective. Their results are open to interpretation, and factors which may influence the listeners' opinions must also be taken into account. This paper investigates one such factor: whether knowing the speaker pairs used in voice transformation affects the listeners' opinions in a subjective listening test concerning the speaker recognizability in voice transformation. This paper also proposes a new type of diagram, called a **Transformation Triangle Diagram (TTD)** to aid in visualizing the results of such a subjective listening test.

2. LISTENING EXPERIMENT DESIGN

Two groups of people, called Group A and Group B, were selected for a voice transformation listening experiment based on the following criteria:

- Each group had 1 pair of male speakers and 1 pair of female speakers.
- When selecting speakers, priority was given to speakers with similar voices based on our subjective opinions.
- The listeners in each group knew the speakers in their group and did not know the speakers in the other group.

For Group A, the female speakers were **clb** and **slt**, and the male speakers were **ehn** and **ref**. For Group B, the female speakers were **hb** and **jm**, and the male speakers were **mo** and **rf**. Each speaker was recorded reading the first 30 sentences of the CMU ARCTIC corpus [3]. Then voice transformation models were trained in both directions for each of the speaker pairs (1 male pair and 1 female pair for each group for a total of 4 pairs). Voice transformation was performed by scaling pitch estimates, using a Gaussian Mixture Model mapping to transform mel-cepstral coefficients, and using a MLSA filter [4] for synthesis as described in [5].

1-4244-0469-X/06/$20.00 ©2006 IEEE

For each speaker pair, a pair comparison evaluation with 10 trials was constructed. The utterances in each pair had different text to avoid confusion from the unmodified portions of source speaker prosody, such as power, that were carried over to the transformed speech. Some trials consisted of recordings from different speakers, some consisted of transformed speech in different directions between the speakers, and some consisted of a recording and transformed speech. The original recordings were analyzed and resynthesized using the same MLSA filter technique [4] employed by the voice transformation process, in order to minimize differences perceived from artifacts due to the vocoding process used during transformation. Listeners were asked to rate the similarity of the speakers in each trial on a scale from 1 to 5, where 1 meant the speakers were very similar and 5 meant the speakers were very different. How the listeners were to judge speaker similarity and difference was left to them. In total, 10 listeners (5 from each group) listened to 40 utterance pairs (10 utterance pairs for each of 4 speaker pairs). With this setup we were able to collect data to investigate whether knowing the speakers made a difference in the judgment of speaker recognizability for voice transformation.

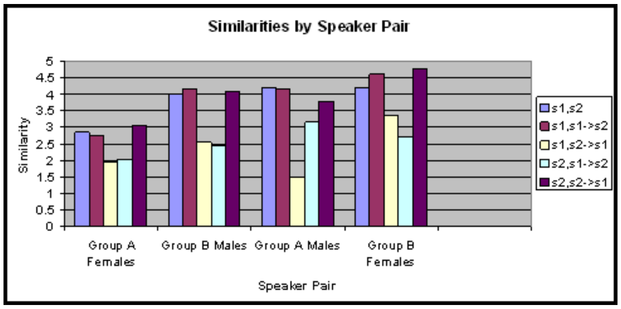

Fig. 1. Similarities by Speaker Pair

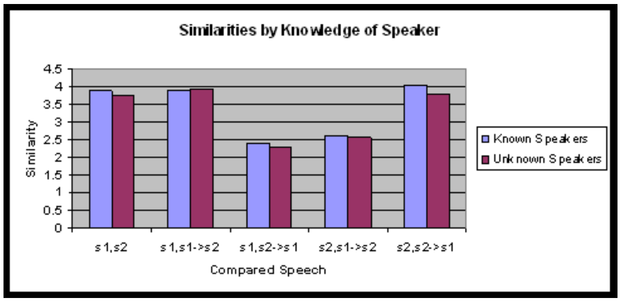

Fig. 2. Similarities by Knowledge of Speaker

3. DATA ANALYSIS

One thing we wanted to know immediately was whether the voice transformation was “successful.” One measure of this was whether the transformed speech was consistently judged as being more similar to the target speaker than the source speaker. This, indeed, was the case when considering the average similarity scores for each speaker pair across all listeners. These averages are shown in Figure 1, where “s1” stands for the first speaker in each pair, “s2” stands for the second speaker in each pair, “s1→s2” stands for transformed speech with the first speaker as the source and the second speaker as the target, and “s2→s1” stands for transformed speech with the second speaker as the source and the first speaker as the target. The scores comparing the target speakers with the transformed speech (s2,s1→s2 and s1,s2→s1) were lower, and thus more similar, than the scores comparing the source speakers with the transformed speech (s1,s1→s2 and s2,s2→s1).

Looking at the bars in Figure 1, a few more trends become apparent. Moving from the leftmost group of bars to the rightmost group, the bars for each speaker pair tend to get higher, showing greater differences in the compared speech. It appears that as the speakers are themselves judged further apart, the transformed speech is also judged as being further from the speakers. The Group A male speakers stand out as having the only exceptions to this general rule. Interestingly, there is a strong asymmetry with the Group A male speakers. The bar comparing the transformation s2→s1 with its target speaker, s1, is much shorter than the bar comparing the transformation s1→s2 with its target speaker, s2. This suggests that the transformation from speaker s2 to s1 was much more successful than the transformation from speaker s1 to s2.

The next question was whether knowing the speakers made a difference. A breakdown of the results according to whether the listeners knew the speakers is given in Figure 2. Not only did the same general trend appear, where the transformed speech was judged as being more similar to the target speech than the source speech, but the scores for each type of compared speech were very close regardless of whether the listeners knew the speakers.

4. TRANSFORMATION TRIANGLE DIAGRAMS

As we looked at numerous graphs similar to the ones in Figure 1 and Figure 2, we realized that we wanted a better way to summarize multiple bars in the graphs and show how their values were related to each other. This led us to create **Transformation Triangle Diagrams (TTDs)** for each speaker pair. Some examples of these are in Figure 3, Figure 4, Figure 5, and Figure 6. TTDs can be interpreted as follows:

- The numbers in the diagrams are calculated by subtracting 1 from the similarity scores to compute 0-based similarity “distances” where 0 is most similar and 4 is most different.

- The distance between speech from the two speakers in a pair is represented by a horizontal line, with the names of the speakers listed at either end.

- Each diagram is composed of two directed triangles.

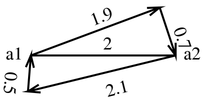

Fig. 3. Transformation Triangle Diagram Example 1

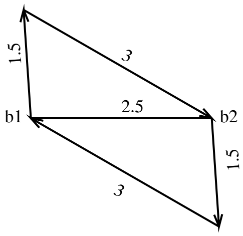

Fig. 4. Transformation Triangle Diagram Example 2

  The upper triangle represents comparisons made using the left speaker in the TTD as the source for voice transformation and the right speaker as the target. The lower triangle represents comparisons made using the right speaker as the source for voice transformation and the left speaker as the target. The arrows serve as reminders for the directions of the transformations.

- The vertices that are off the horizontal baseline represent transformed speech, and the remaining triangle edges represent the distances from the speakers' speech to the transformed speech. For example, in the first TTD in Figure 3, the distance between speaker a1 and speech transformed from a1 to a2 is 1.9, the distance between speech transformed from a1 to a2 and speaker a2 is 0.7, the distance between speaker a2 and speech transformed from a2 to a1 is 2.1, and the distance between speech transformed from a2 to a1 and speaker a1 is 0.5
- It should be noted that TTDs make no attempt to compare transformed speech using one speaker as the source with transformed speech using the other speaker as the source.

A few examples of TTDs are given in Figure 3, Figure 4, Figure 5, and Figure 6. Figure 3 represents a pair of speakers called a1 and a2, where both transformations were mostly successful in that the transformed speech was considerably closer to the targets than the sources in both cases.

Figure 4 represents a pair of speakers called b1 and b2, where both transformations were fairly unsuccessful in that the transformed speech was closer to the source than the target. As transformation becomes more successful, the TTDs tend to skew so the upper triangle is crushed to the right and the lower triangle is crushed to the left.

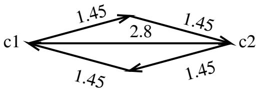

Fig. 5. Transformation Triangle Diagram Example 3

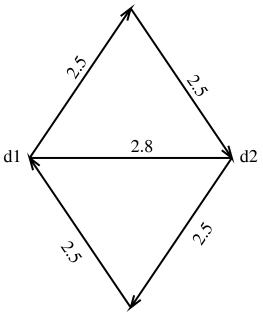

Fig. 6. Transformation Triangle Diagram Example 4

However, distance from a vertex representing transformed speech to the horizontal baseline can make a difference as well. In Figure 5 representing speakers c1 and c2 and in Figure 6 representing speakers d1 and d2, the vertices representing the transformed speech would project to the same location on the horizontal baselines, but the transformations between c1 and c2 were more successful than the ones between d1 and d2 because the transformed speech is closer to the targets. One additional point is that the length of the horizontal baselines vary according to the similarity of the speakers. The more similar the speakers are, the narrower the baseline is.

In the ideal case, both transformations would coincide with their targets, and the TTD would collapse to a horizontal line with arrowheads pointing outward at each end. In a case where the transformation was completely unsuccessful and the transformed speech sounded like the source voice, the TTD would again collapse to a horizontal line, but there would be inward pointing arrows as well.

It is important to note that the distances in these diagrams may not actually be distances in a Euclidean sense, and it may not be possible to construct triangles for some combinations of scores if the lengths of the edges do not satisfy the triangle inequality. One pathological case would be when the horizontal bar is longer than the sum of the other two sides of a triangle. That would mean that the distance between the source and target speakers is actually greater than the combined distances of the transformed speech to both the source and target speakers. The other pathological case would be when the distance from the transformed speech to one of the speakers was greater than the sum of the distance from the transformed speech to the other speaker plus the distance between the two speakers themselves. In such a case, it would also be impossible to construct a triangle. However, it should be noted that for all the examples we tried based on our data, we were able to construct triangles

TTDs are not the first attempt to try to represent distances between speech in voice transformation. Others have used Multi-Dimensional Scaling (MDS) techniques to accomplish this [6]. In MDS, distances are calculated among multiple

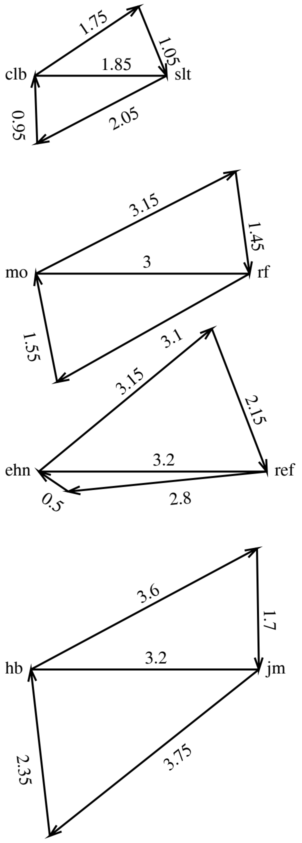

Fig. 7. Transformation Triangle Diagrams by Speaker Pair

quantities in a multi-dimensional space, and the results are projected onto a plane for comparison. Although MDS is an interesting and useful technique for analyzing data, we find that TTDs are a compact, simpler-to-understand way of depicting the specific relationships we are trying to compare in voice transformation.

5. EVALUATING VOICE TRANSFORMATION WITH TTDS

The TTDs for results from our listening experiment broken down by speaker pair are in Figure 7. These results correspond to the four speaker pairs from the graph in Figure 1. Looking at these TTDs, a number of things become readily apparent. First of all, the transformations were mostly successful in the sense that the triangles are skewed so the transformed speech is closer to the target speech than the source speech in each case. Another point is that the speakers in the first pair were considered much more similar than the others based on the widths of the diagrams. One interesting thing that appears in the third pair is that the transformation from **ref** to **ehn** is much more successful than the transformation from **ehn** to **ref**, as shown by the asymmetry in the diagram. This is another visual depiction of the same asymmetry mentioned earlier in the section on Data Analysis.

6. CONCLUSIONS

In our listening experiment, we found that whether the listeners knew the speakers did not appear to significantly affect how they judged speaker similarity. This knowledge will guide us in designing further experiments of this nature because we will not be concerned with finding listeners who either know or don't know the speakers. We have also created a new type of diagram called a **Transformation Triangle Diagram (TDD)** that was useful in representing certain relationships in a compact, understandable manner. Future work will involve investigating further methods of visualizing voice transformation results. While this paper investigates the area of speaker recognizability, there are other areas of voice transformation evaluation, such as intelligibility and naturalness, where different forms of analysis may be necessary.

7. ACKNOWLEDGMENTS

This work was supported by the US National Science Foundation under grant number 00414675 "TRANSFORM: flexible voice synthesis through articulatory voice transformation". Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

8. REFERENCES

[1] Tomoki Toda and Kiyohiro Shikano, "NAM-to-speech conversion with gaussian mixture models," in *Interspeech 2005*, 2005, pp. 1957–1960.

[2] A. Kain, *High Resolution Voice Transformation*, Ph.D. thesis, OGI School of Science and Engineering, Oregon Health and Science University, 2001.

[3] J. Kominek and Black A., "The CMU ARCTIC speech databases for speech synthesis research," Tech. Rep. CMU-LTI-03-177 http://festvox.org/cmu_arctic/, Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, 2003.

[4] S. Imai, "Cepstral analysis synthesis on the mel frequency scale," in *Proceedings of ICASSP 83*, Boston, MA, 1983, pp. 93–96.

[5] Tomoki Toda, *High-Quality and Flexible Speech Synthesis with Segment Selection and Voice Conversion*, Ph.D. thesis, Nara Institute of Science and Technology, 2003.

[6] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," in *Proceedings of ICASSP 88*, Tokyo, 1988, pp. 655–658.