CSCE 681 600: SEMINAR REPORT 1

Name: Rohan Chaudhury

UIN: 432001358

Paper Citation: Felps, Daniel; Gutierrez-Osuna, Ricardo, "Developing Objective Measures of Foreign-Accent Conversion," in IEEE Transactions on Audio, Speech, and Language Processing, vol. 18, no. 5, pp. 1030-1040, July 2010, doi: 10.1109/TASL.2009.2038818.

1. SUMMARY:

a. Problem Statement:

The authors in this paper have presented their work based on the conversion of foreign accented speech to native accented speech. Generally, accent conversion means the generation of new utterances from the original utterances of the foreign speaker but in a native accent. Their main aim in this work was to devise efficient objective measures to assess various accent conversion methods that have recently been developed so that they may be used to improve the efficiency of the newly developed accent conversion methods. The authors have identified 3 objective measures to evaluate the acoustic quality, foreign accent degree, and the identity of the speaker in the speech after accent conversion. The authors have also shown an accent conversion method in this paper which is based on the notion that the foreign accent in the original utterances can be removed by applying convolution on the extracted voice quality carrier from the original foreign accented speech with the linguistic gestures in a native accented speech. They have proposed that the objective measures identified by them can also be used in computer assisted pronunciation training (CAPT) to provide an assessment to the L2 learner on their speech and accent and also to find an equivalent voice from a set of native speakers which complements the voice of the learner.

b. Proposed solution:

The authors have first described the methods that they have used for accent conversion. They have first done prosodic conversion which is achieved by changing the phoneme durations and the contour of the pitch of the original utterance in the foreign accent into those of the target in the native accent. They have assumed that the speech is segmented phonetically and from these segments, they have identified the source to target duration ratio to get the time-scaling factor alpha in order to perform time-scaling on the source utterances on each phoneme basis. For pitch scaling, they have replaced the contour of the pitch of the original source utterance with a scaled and shifted version of the contour of the pitch of the target native utterance. They have limited the pitch scaling by a factor beta. By maintaining the range and baseline of the pitch the authors were able to conserve the identity of the speaker.

Then they have performed segmental accent conversion where they have combined the glottal excitation of the source utterances with the spectral envelope of the target. They have done so because the authors had assumed that the voice quality is dependent on signals of glottal excitation and the linguistic content is dependent on the filter (spectral envelope in this case). After doing segmental accent conversion they had obtained a signal which was a combination of the excitation of the source and spectral envelope of the target which was normalized to the vocal tract length of the original source.

Then they have performed perceptual experiments to determine the acoustic quality decline, reduction of foreign accent degree, and how much the identity of the original speaker was conserved. For acoustic quality decline, participants gave MOS value ratings to the acoustic quality of the utterances. For reduction of foreign accent degree, participants gave ratings on a seven-point Empirically Grounded, Well Anchored EGWA scale to the degree of foreign accent in the utterances. For speaker identity, the participants, after listening to two different utterances, were asked to rate their confidences on the EGWA scale about whether the utterances belonged to the same speaker or not.

Then they explained the objective measures that they defined to assess the accent conversion systems. To assess acoustic quality, they used an objective measure which was based on the recommendation P.563 by ITU-T for single-ended speech quality. Preprocessing, estimation of distortion and perceptual mapping were the 3 stages used in their algorithm.

To assess foreign accent reduction degree, they had evaluated the test utterances using a continuous speech Hidden Markov Model which was trained on acoustic models from native speakers. The score obtained from the model gave an estimate of the degree of nativeness.

To assess the speaker's identity, the authors had used Fisher's Linear Discriminant Analysis (LDA). By using LDA they were finding a projection that maximized the separability between 2 given source and target speakers based on a given corpus of acoustic features from both the speakers. Based on the average LDA projection of the frames of the test utterances, it is given an identity score which helped determine the identity based on the closeness from either the scores of the source or the target.

c. Results:

The authors had used a 2 way analysis of variance (ANOVA) to calculate the statistical significance based on the prosodic and segmental transformation factors. In the case of acoustic quality, the participants found detectable distortions in the output after prosodic and segmental conversion modifications. Also, they found that the quality of source recordings were higher than that of target. This supported their hypothesis that lower scores were given by participants due to the lower intelligibility of foreign accents and not because of the recording quality.

In the case of measuring foreign accent degree, the authors observed that the effect of interaction and prosody effects were not that significant whereas the effect of segmental transformation was significant in decreasing the foreign accent degree. For the unexpected results with respect to prosody and interaction effects, the authors gave a possible explanation that the prosody of the source and the target might have been similar from the start.

In the case of speaker identification, it was observed that the participants were able to properly differentiate between the source and segmental converted target. However, they were not able to differentiate between the source and prosodic converted target or between segmental converted target and segmental+prosodic converted target.

2. CRITIQUE:

a. Pros:

i. The authors had done thorough research on past work done in this field. This allowed them to identify the various pros and cons of previous research and use them accordingly in their paper.

ii. The authors have identified very useful objective measures which can efficiently assess accent conversion methods as demonstrated by their results. Their objective measures will help advance the development of these methods in the future.
iii. Objective evaluation by these objective measures can potentially replace the perceptual evaluation of accent conversion methods which involves manual labor.
iv. Objective measures can also be used in CAPT so that learners can learn the language on their own by self-evaluating themselves from the feedback received.

b. Cons:

i. Only 3 objective measures were identified to assess the accent conversion methods. There could be many more objective methods that can be used to do the evaluation.
ii. Limited methods were used to assess the 3 objective measures. In the case of Foreign accent analysis, instead of HMM, deep neural networks such as Recurrent Neural Networks or Transformers could be used.
iii. Deep neural networks could also have been used in the case of identifying the speaker.
iv. The participants were presented with the recordings played backward during perceptual evaluation to identify the speaker. This removed the vocabulary, language, and accent which can be a very important factor to identify speakers.
v. A more thorough perceptual evaluation with an increased number of experiments was needed to be done.

3. FOLLOW UP:

i. Research on identifying more objective measures to assess the accent conversion methods can be done for their proper assessment.
ii. Deep Neural Networks can be explored in some cases of the assessment, such as RNNs or Transformers can be used for foreign accent analysis or speaker identification.
iii. Other methods of scoring the objective measure values can be identified.
iv. The ethnicity of an individual heavily influences the accent of an individual, with this assumption research can be done to identify the ethnicity of an individual from their utterances.
v. Accent conversion methods based on different ethnic accents can be developed which will give more accurate accent conversions.
vi. Accent conversion methods using RNNs and Transformers can also be explored.