PERTURBATION AND PITCH NORMALIZATION AS ENHANCEMENTS TO SPEAKER RECOGNITION

A. Lawson[1], M. Linderman[2], M. Leonard[3], A. Stauffer[1], B. Pokines[4], M. Carlin[2]

[1]RADC, Inc.
[2]Air Force Research Laboratory
[3]University of Texas at Dallas
[4]Oasis Systems, Inc.

aaron.lawson.ctr@rl.af.mil, mwl53@cornell.edu, mrl016000@utdallas.edu, stauffar@clarkson.edu, Benjamin.Pokines.ctr@rl.af.mil, macarlin@jhu.edu

ABSTRACT

This study proposes an approach to improving speaker recognition through the process of minute vocal tract length perturbation of training files, coupled with pitch normalization for both train and test data. The notion of perturbation as a method for improving the robustness of training data for supervised classification is taken from the field of optical character recognition, where distorting characters within a certain range has shown strong improvements across disparate conditions. This paper demonstrates that acoustic perturbation, in this case analysis, distortion, and resynthesis of vocal tract length for a given speaker, significantly improves speaker recognition when the resulting files are used to augment or replace the training data. A pitch length normalization technique is also discussed, which is combined with perturbation to improve open-set speaker recognition from an EER of 20% to 6.7%.

***Index Terms*—s**peaker recognition, speech synthesis

1. INTRODUCTION

Data perturbation and perturbed synthesis has been demonstrated to mitigate the effects of train and test data mismatch in the field of Optical Character Recognition (OCR). In OCR, as in speech, it is often possible to obtain a large amount of high quality training samples for a given phenomena, but very difficult to anticipate the effects of the real-world, sub-optimal conditions under which the system will be required to function. Often, in both OCR and speech processing, training data of sufficient quantity and which adequately reflects natural distortion is simply not available. This is especially true for speaker identification, where one may have only a single sample of a target's voice. In OCR, the process of perturbation is applied to data that is too clean, too uniform, or too reflective of a particular set of conditions, such as a single typeface or a small set of handwriting samples. Perturbation [1] [2] is applied to training samples to minutely distort the data such that the characters are still recognizable as themselves, but are less uniform, clear or clean. Perturbation may be applied directly to the training data or the training data may be synthesized with perturbation for digital data. The end result in OCR is a major improvement in the robustness, accuracy and generalizability of character models.

The current study applies the concept of perturbation to speaker recognition. More specifically, we explore the automatic augmentation and enhancement of training data for a given set of speakers using vocal tract length perturbation within a predetermined range. This paper also examines another acoustic data manipulation technique of pitch normalization. Both these phenomena are generated during resynthesis using a speech analysis, rescaling and synthesis system based loosely on concepts from speech manipulation systems such as the STRAIGHT algorithm [3]. The immediate goal of this approach is to counter the effects of session variability and the natural small shifts in vocal register that occur during conversation due to mood, speaker interaction factors, changes in the noise environment, etc.

2. ACOUSTIC PERTURBATION APPROACH

In the context of this paper acoustic perturbation will refer to the analysis and rescaling of vocal tract length (VTL), which will then be used to synthesize a new, slightly distorted, speech sample. The tool used to perform this analysis and synthesis is a C-code suite of tools based on the ideas of the STRAIGHT algorithm. An optimized C version of this tool was important in order to render the required transformations in real-time, due to the complexity of the analysis and synthesis process.

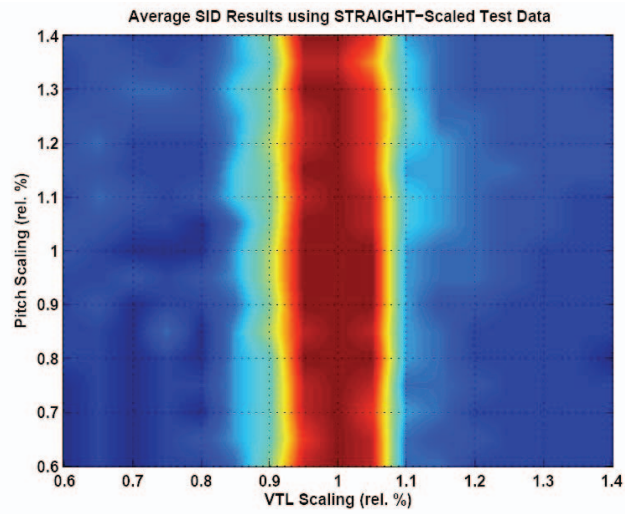
Initial experiments were performed to determine the acceptable range of VTL rescaling, such that the new sample is different from the original, but is still a sample of the target speaker. Speakers from the TIMIT corpus were rescaled on a range of values from 60% to 140% of their recognized VTL, where the percentage is relative to the speaker's VTL as determined by the analysis algorithm. Each scaled set was then evaluated with a speaker ID system to determine the range where altering VTL results in rejection of target, and the range where said speaker is safely IDed as him/herself. Based on these tests it was determined that scaling VTL beyond +/- 15% will safely create a novel speaker, and that rescaling within +/- 7% will generate the same target.

3. PITCH NORMALIZATION APPROACH

While pitch is largely a function of excitation (F0), changes in pitch do impact the nature of higher formants, with higher pitch rendering more widely spaced formants and lower pitch

compressing the formant space. For lower sample rate recordings, female speakers with high F0 may not exhibit the F4 formant, or

Chart 1: Impact of VTL and Pitch rescaling on speaker Recognition



the F4 formant may be very exceedingly weak or absent in many vowels. Studies have shown that both F3 and F4 are important to speaker recognition [4] and that strengthening the presence of higher formants can improve the robustness of cross-session speaker recognition [4]. Moreover, while vocal register differences tend to manifest themselves in F0, they also impact the shape of formants in the higher frequencies [5]. Even in normal conversation, F0 changes widely, with a typical range for males varying between 211 and 86 Hz , due to differences in interactive situation, lexical tone, and sentence type (interrogative vs declarative, etc. [6] Research has shown that such changes in intonation impact speaker recognition [7]. Normalizing pitch is hypothesized to partially eliminate these distracting artifacts so that stable, speaker specific features may be more readily extracted.

4. EXPERIMENTAL PROCEDURE

Open set experiments were run on two corpora to determine the impact of perturbation and pitch normalization on speaker recognition. The major hypotheses were

1) Does VTL-based acoustic perturbation have an impact on speaker models? If so, what scalings are optimal?

2) Is it more effective to combine synthesized data with original data in the same model or by building separate models?

3) Does pitch normalization impact speaker recognition? For both training and test data? For which pitch ranges?

4) Can perturbation and pitch normalization be combined to further improve speaker recognition results?

4.1 The Speaker Recognition System

The Gaussian Mixture Model (GMM) and Universal Background Model (UBM) approach, developed by Reynolds [8], are also used in this study. Front-end feature processing consists of mel-weighted and delta fft-cepstra generated from a frame size of 20ms with 50% overlap. During recognition, the likelihood of the test speech is computed for each of the GMMs produced during training. For the implementation used in this paper only 5 mixtures are used for the calculation of the likelihood of a particular speaker's GMM model. The five mixtures are chosen from the most probable mixtures in the UBM. The goal of this paper clearly does not lie in altering the speaker recognition system, or of comparing its accuracy to other algorithms, but rather in evaluating the impact of the two experimental conditions on how a speaker recognition system performs.

4.2. Databases

For the purposes of this study two corpora were chosen, with different channel and session characteristics. For same-session tests the CSLU 22 Language Corpus, a monologue telephone database, was used. Ninety speakers total were involved, all English speakers with 50 seconds of continuous speech. Thirty speakers were used as targets in each test run, with 60 as impostors. Each training sample was 25 seconds of speech, as was each test sample. For inter-session tests the Multi-Session Audio Research Corpus (MARP) [7] was used, as it contained speech data spanning 21 sessions over three years, and could thus be used to verify performance across sessions. For these initial experiments only pitch normalization was run on the MARP data, since test and train consisted of single sentences (0.4-1.7 sec. in length) and effective VTL scaling could not be accomplished. MARP tests were round-robined with 17 targets and 17 imposters. Experiments are continuing with the conversational portion of MARP.

4.3 Experimental Parameters

Acoustic perturbation varied along three parameters: 1) the degree of scaling performed, 2) the composition of models in terms of combinations of different scalings, 3) whether a single model was created for each speaker or if multiple models were used. Pitch normalization varied across two parameters: 1) the pitch to which speakers were scaled (62.5 Hz, 125 Hz, and 250 Hz) and 2) whether just training data was normalized, or if both train and test were normalized. A final set of experiments combined both perturbation and pitch normalization.

5. RESULTS

Tests were run using the following configurations: 1) "Single Models" are created by combining all the training data into a single model for each speaker. For example, both pitch normalized and original data would be combined to build a single model for a given speaker. 2) For the "Multi-Model" approach several different models are generated for each speaker, each representing a different synthesis condition, often along with a model for the original data.

5.1 Perturbation

A "stacked" testing approach was used to produce perturbation results where only the original file was used to train a model to start, then the unaltered synthesized file was added to the model (100s, for "100% scaling") and tested, followed by plus and minus 2% VTL files, 4% VTL files, and so on. The combination of synthesized files from 96-104% VTL produced the most effective model and reduced EER from 0.194 to 0.092, as seen in Table 1.

Table1: Results of Perturbation Parameters

Training Data	EER	
Parameters	*Multi-Model*	*Single Model*
Original only	0.194	0.194
Original & 100s	0.158	0.192
98-102s	0.129	0.138
96-104s	0.100	0.092
94-106s	0.100	0.125
92-108s	0.100	0.142
90-110s	0.087	0.133
88-112s	0.100	0.167

5.2 Pitch Normalization

Experiments with pitch normalization included three F0 settings of 62.5Hz, 125Hz, and 250Hz. The three modeling approaches included the single model and multi model for both real and synthesized data and another single model approach for synthesized data only. Results indicate normalization to an F0 of 125Hz to be most effective. This setting affords a realistic pitch setting within the typical male range of 86-211 Hz and optimizes for 8k sampling rate audio.

Table 2: Pitch Normalization Results -Test Pitch Not Normalized

Train Pitch	Model Type	EER
	Baseline	0.194
62.5Hz	Single-model	0.158
	Multi-model	0.158
	Synth only	0.213
125Hz	Single-model	0.133
	Multi-model	0.150
	Synth only	0.267
250Hz	Single-model	0.200
	Multi-model	0.196
	Synth only	0.329

Moreover, viewing of the spectrogram shows a loss of formant structure at 62.5Hz and an information "undersampling" at 250Hz. Nevertheless, the 62.5 Hz F0 still provides improvements over non-normalized audio, though slightly less than the optimal 125Hz frequency. Table 2 shows an improvement in EER from the baseline of 0.194 to 0.133 with a single, composite model pitch normalized at 125 Hz. With the test data normalized as well we see multi-model and pure synthesized data reaching an EER of 0.133 (Table 3) and error rates for synthesized data dropping significantly in every category. Table 4 gives results for the MARP cross session data, where greatest reduction in EER was obtained at 125 Hz using the synthesized data sets, with a 4.4% absolute reduction in EER.

5.3 Combined Pitch Normalization and Perturbation

Table 5 represents results where both pitch normalization and VTL scaling perturbation are applied to the training data files, with test files left unchanged. Only the 96-104 scaling was tested as it was the best performer in the perturbation-only set of tests.

Table 3: Pitch Normalization Results -Test Pitch Normalized (CSLU)

Train Pitch	Model Type	EER
	Baseline	0.194
62.5Hz	Single-model	0.158
	Multi-model	0.200
	Synth only	0.150
125Hz	Single-model	0.167
	Multi-model	0.133
	Synth only	0.133
250Hz	Single-model	0.142
	Multi-model	0.200
	Synth only	0.233

Table 4: Pitch Normalization Results -Test Pitch Normalized (MARP-Cross-Session)

Train Pitch	Model Type	EER
	Baseline	0.301
62.5Hz	Single-model	0.289
	Multi-model	0.341
	Synth only	0.297
125Hz	Single-model	0.262
	Multi-model	0.318
	Synth only	0.256
250Hz	Single-model	0.358
	Multi-model	0.398
	Synth only	0.370

Table 5: Combined Perturbation and Pitch Normalization -Test Pitch Not Normalized (CSLU)

Train Pitch	Test Pitch	Model Type	Perturbation	EER
62.5Hz	Actual	Single Model	96-104s (pitch synth with real)	0.100
			96-104s synth only	0.204
		Multi-model	96-104s (real and synth)	0.133
			96-104s synth only	0.217
125Hz	Actual	Single Model	96-104s (pitch synth with real)	0.133
			96-104s synth only	0.200
		Multi-model	96-104s (real and synth)	0.129
			96-104s synth only	0.196

This combination with pitch normalization of train files only, provides error reduction to 0.100 EER, a level previously achieved with perturbation alone.

Table 6: Combined Perturbation and Pitch Normalization -Test Pitch Normalized

Train Pitch	Test Pitch	Model Type	Perturbations	EER
62.5Hz	62.5Hz	Single Model	96-104s (pitch synth with real)	0.100
			96-104s synth only	0.096
		Multi-model	96-104s (real and synth)	0.100
			96-104s synth only	0.096
125Hz	125Hz	Single Model	96-104s (pitch synth with real)	0.104
			96-104s synth only	0.096
		Multi-model	96-104s (real and synth)	0.067
			96-104s synth only	0.067

However, in the matched case (Table 6) when the test pitch is normalized to match the train, error rates are further reduced from a previous best of 0.092 to 0.067 EER. Overall this combination produces a notable improvement from a baseline equal error rate of 19.4% to 6.7% and demonstrates the utility of combining both perturbation and pitch normalization within SID systems.

6. DISCUSSION/CONCLUSIONS

To return to the four research questions posed in section 4, we find that VTL-based acoustic perturbation has a beneficial impact on the accuracy of speaker models, with a decrease in EER from 0.194 to 0.092, with the scaling combination range of 96-104% VTL having the best effect.

Pitch normalization on its own did reduce EER as well, though not as much, with a drop from 0.194 to 0.133 on CSLU and a drop from 0.301 to 0.256 on the MARP data. Matching pitch in test and train did not impact lowest EER in the pitch normalization-only experiments, but it did reduce the amount of error in tests overall and it had a significant impact in the combined pitch/perturbation tests. Best performance was achieved at a scaling of 125 Hz in almost all cases. Results clearly show that combining pitch normalization and VTL perturbation decreases EER in these experiments, with combined performance reducing error from the baseline error of 0.194 to 0.067, and from the error rate of VTL perturbation alone at 0.092, down to to 0.067.

The notion that speaker data reflecting a wider range of vocal conditions would improve speaker recognition is intuitive and well supported by experimentation. Of interest here is the use of automatically synthesized data, albeit transformed rather than generated, to improve speaker identification models. The authors' initial impression that synthesis was too crude to be of use in target identification was changed by studies showing that language identification (LID) performance could be improved by using STRAIGHT to generate novel speakers for augmentation of a LID model [9]. Furthermore, pilot experimentation showed that moderately vocoded data could augment speaker identification models for use on non-vocoded test data. This demonstrated that while human intelligibility may suffer in the process of speech transformation, information useful to speaker identification is often maintained. This is analogous to the loss of information in the use of RASTA filters, which can also serve to improve the generalizability of speaker recognition.

The impetus of this study came out of a realization that speaker voice changes even over the short term (within a single conversation) can dramatically impact speaker recognition, and that approaches to mitigating this effect are essential to combating the larger problem of inter-session variability. Most research has focused on the accompanying interference that impacts speaker recognition across sessions, such as channel differences, noise, and digital distortion. What has received much less attention has been the variation in the voice itself, and how the voice changes in reaction to noise, channel levels, distortion, environment, interlocutor, emotion, etc. This paper presents one approach to data enhancement using the voice characteristics of pitch and vocal tract length to provide a way of expanding the recognition power of a model built with a small sample of audio to greatly reduce error rates.

8. REFERENCES

[1] T.M. Ha and H. Bunke, "Handwritten Numeral Recognition by Perturbation Method," *Proc. Fourth Int'l Workshop Frontiers of Handwriting Recognition,* Taipei, Taiwan, pp. 97-106, 1994.

[2] Yasuda, M., K. Yamamoto, and H. Yamada, "Effect of the perturbed correlation method for optical character recognition," *Pattern Recognition,* Volume 30, Issue 8, pp. 1315-1320, August 1997.

[3] H. Kawahara and H. Matsui, "Auditory Morphing Based On An Elastic Perceptual Distance Metric In An Interference-Free Time-Frequency Representation", *Proc. ICASSP 2003*, vol. I, pp.256-259, 2003.

[4] Marrero, V. et al. "Identifying speaker-dependent acoustic parameters in Spanish vowels" *Proceedings of Acoustics '08, Acoustical Society of America*, Paris, France, 2008.

[5] M. Zemke, I. Tokuda, and H. Herzel, "Modeling of voice registers and bifurcation theory" " *Proceedings of Acoustics '08, Acoustical Society of America*, Paris, France, 2008.

[6] Wang, M, "An Analysis of Pitch in Chinese Spontaneous Speech", *ISCA International Symposium on Tonal Aspects of Languages*, Beijing, China, 2004.

[7] Lawson, A., Stauffer, A, Wenndt, S., "External factors impacting the performance of speaker identification in the Multisession audio research project (MARP) corpus", *153rd Meeting of the Acoustical Society of America*, June 4-8, 2007

[8] Reynolds, D. A. "Comparison of Background Normalization Methods for Text-Independent Speaker Verification." *Proceedings of the European Conference on Speech Communication and Technology*, Rhodes, Greece, Vol. 2, pp. 963-966, 1997.

[9] A. Lawson, M. Linderman, M. Carlin and A. Stauffer, "Automatic data enhancement for language identification using voice generation", *Proceedings of Acoustics '08, Acoustical Society of America*, Paris, France, 2008.