MFCC FEATURE EXTRACTION IMPROVEMENT FOR NOISY ENVIRONMENTS WITH GENDER MISMATCH: DYNAMIC ALPHA DETERMINATION FOR VTLN AND SINGLE CHANNEL NOISE REDUCTION WITH WIENER-SCARLART FILTER

Miles Abel, Nicholas Ma

Department of Electrical Engineering, University of California Los Angeles, CA 90095, USA

ABSTRACT

This paper proposes a new way to dynamically determine the warping scale factor, alpha, for Vocal Tract Length Normalization (VTLN) and explores lesser known noise reduction algorithms to improve an isolated-word recognizer. It addresses speech feature extraction improvements in two areas: gender mismatch and noisy environments. For gender mismatch, we propose dynamically determining the scale factor based on the probability distribution of female pitch. For noise robustness, this paper will present a Single-Channel Noise Reduction (SCNR) algorithm based on a Magnitude-DFT (MDFT) estimator in conjunction with a complex-DFT estimator (CDFT) to reduce the effect of noise on feature extraction for speech [1].

1. INTRODUCTION

The objective of this paper is to discuss methods of generating a robust isolated-word recognizer of connected speech, operating within the environmental constraints of noise and gender mismatch. Specifically, we seek to explore new and less commonly used methods for feature extraction that improve speech recognition in a speech processing system using an HTK hidden Markov model (HMM) that uses Mel-frequency cepstral coefficients (MFCC) feature extraction as the baseline. Because we address two independent constraints, gender and noise, the paper will cover each separately. Machine learning models typically train on male speech data, as does the system in this report. As a result, female speech often becomes unrecognizable by a male trained HMM based machine learning algorithm. Speech also exists in the presence of noise. When noise is present in the same frequency spectrum as speech, our ability to decipher the contents of a speech sample decrease substantially.

At present, automatic speech recognition (ASR) is an area of great interest due to the convergence of new research in machine learning methods, such as deep learning. In addition, there are growing capabilities and demands in several different industries. This makes robust speech enhancement not only a standalone human achievement but also an area of significance in the ever-developing world of artificial intelligence (AI), robotics and automation.

2. BACKGROUND

The issue of unknown gender gives rise to challenges originating from differences in the vocal tract lengths of men and women. Men on average have a vocal tract length of 16.9 cm, whereas the average woman has a length of 14.1 cm [2]. Pitch is directly affected by the length of the vocal tract with the average male pitch slightly above 100 Hz and the average female pitch a little above 200 Hz [3]. However, adding to the difficulty, the pitch range overlaps between the two genders. A well-established method for accounting for the differences in pitch is Vocal Tract Length Normalization (VTLN). VTLN stretches or compresses speech signals to warp the pitch of a given signal to a reference pitch [4]. Common warping functions are symmetric, asymmetric, quadratic, power and bilinear [4]. For each method, there is a scaling factor, alpha. However, this warping scaling factor is constant throughout these methods. Another method implementing VTLN is to warp the center frequencies of a Mel filter-bank and use the maximum likelihood score (MLS) to estimate the warping factor [5]. The focus for handling the issue of gender differences in speech will be novel approaches to warping the frequency of female speech by gleaning information from the female pitch distribution.

Machine learning algorithms, such as the HTK HMM used in this report, build models of unknown states and assign probabilities to state transitions and outcomes from each state based on the data provided for training the model [6]. Typically, speech training data is clean, i.e. without noise, to allow the HMM to build a model that represents actual words. However, the model encounters noise, of a varying and unpredictable nature, in almost all scenarios of practical application. Thus, to design a robust isolated-word recognizer, feature extraction should be void of noise for the HTK HMM to accurately determine the words spoken. We investigate several of the less commonly used methods for removing noise in feature extraction.

In recent years, an introduction to Power Normalized Cepstrum Coefficients (PNCC) has shown great improvements in terms of feature extraction from noisy speech. PNCC processing uses power-law nonlinearity to replace the log nonlinearity commonly used in MFCC. The algorithm uses asymmetric filtering that suppresses background excitation and temporal masking. Medium power analysis and frequency smoothing further allows suppression of unwanted noise. Although PNCC does an excellent job of noise reduction, the errors between training male speech and testing female speech becomes more apparent from the MFCC baseline test.

Regarding noise reduction algorithms, one of the earliest models is spectral subtraction. Spectral subtraction estimates the noise spectrum during speech pause and subtracts it from the speech spectrum. The drawback of this method is the presence of processing distortions - remnant noise [7].

Given the problems with gender mismatch using PNCC as well as remnant noise using spectral subtraction, we explore various noise

reduction algorithms that can both maintain the baseline gender mismatch and further reduce the noise of the signal. From this, we investigate two different Discrete Fourier Transform (DFT) based estimators where one uses the magnitude-DFT (MDFT) and the other uses the complex-DFT. Hendriks' Single Channel Noise Reduction (SCNR) algorithm incorporates the MDFT based estimator whereas the Wiener-Scarlart algorithm uses CDFT based estimator.

3. PROJECT DESCRIPTION

3.1. Gender Mismatch

We explore ways for dynamically determining the alpha value used to warp Mel-filter-bank frequencies in VTLN. Our objective is to determine if the likelihood of a particular female pitch can be used to influence the alpha value. For this report, we use the 500 samples of clean testing female data to generate our female pitch distribution. In an ideal situation, our distribution of female pitch would have an infinite sample size or at least one that contains hundreds of thousands of female speakers. To calculate the pitch, we use the MATLAB script, fast_mbsc_fixedWinlen_tracking.m [8]. With this script, we are able to create a vector containing the pitch per frame for each sample of the clean testing female data. We then remove all 0's and NaN's and calculate a median pitch vector, Equation 1.1.

$$pitch\_Ftst = \begin{bmatrix} \\ \vdots \\ \\ \end{bmatrix} \tag{1.1}$$

In order to gather meaningful information from the pitch, we compute the probability density function (PDF) of the pitch vector in Equation 1.1 using a univariate normal kernel smoothing function in MATLAB called ksdensity(). Figure 1 displays the probability density function for both the clean male training and clean female testing data.

The next challenge we face is determining the range with which to dynamically control alpha. One option would be to dynamically change the alpha value over all frequencies. However, because we train on clean male speech and test on both male and female speech, we choose to only warp the alpha value when the speaker is female. Then, the problem of determining where female pitch begins and male ends arise as can be seen from the overlap of the blue male PDF and red female PDF of pitch in Figure 1. There is no exact point where female pitch ends and male pitch begins but rather we seek to find an approximate transition point for the pitch based on the probability density distributions.

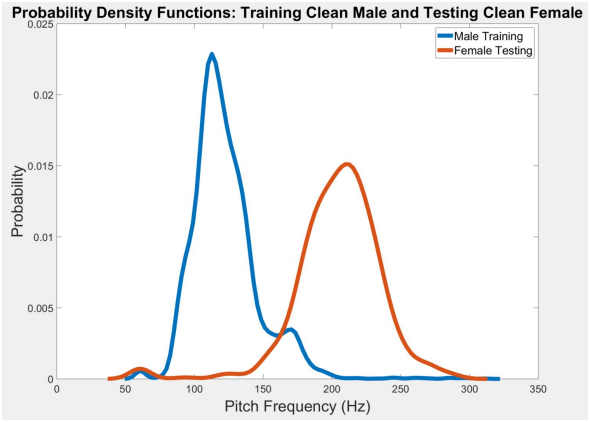

Fig. 1. PDF of both the median pitch vector for the training clean male data and testing clean female data

Initially, we began with constant alpha values and use them to scale the center frequency of the Mel-filter-bank. The alpha values are constant scale factors for the center frequencies of the Mel-filter-bank. In Table 1, the results of several different combinations of boundary conditions and constant alpha values are shown.

Constant Alpha Clean Female Calculations			
Flpitch	Fhpitch	alpha	Clean Female
0	∞	1	82.72%
0	∞	1.1	97.82%
0	∞	1.05	95.06%
0	∞	1.15	95.77%
0	∞	1.105	98.30%
0	∞	1.11	97.71%
0	∞	1.09	97.00%
0	∞	1.095	97.35%
0	∞	1.099	97.59%

Table 1. Various constant alpha values for VTLN. Flpitch is the lowest pitch at which alpha is applied. Fhpitch is the upper frequency limit of pitch, which has no limit.

Although the results of using a constant alpha value are favorable, a constant alpha value does not allow us to dynamically adjust its value based on the probability of a female pitch. We have a maximum accuracy of 98.30% with an alpha value of 1.105, which is an increase in accuracy of 15.58% over the baseline of 82.72%. However, we aim to use the probability density distribution of female pitch to modify alpha and continue with a linear alpha relationship.

In Table 2, we have various linear alpha versus pitch frequency relationships. These are all piecewise functions with constant alpha values below and above specific thresholds and are linear in between these boundary conditions. All the results improve the performance of the clean female testing data compared to the 82.72% baseline, which has an alpha value of 1 in Table 1. For this linear case, the choice of boundary conditions allows for the probability distribution of female pitch to influence the alpha value to some degree.

Linear Alpha Clean Female Calculations						
Flpitch	Fhpitch	slope	intercept	alphaL	alphaH	Clean Female
195	300	1/525	22/35	1	1.2	92.42%
195	350	1/775	116/155	1	1.2	90.36%
150	300	1/750	4/5	1	1.2	96.94%
150	300	1/857	33/40	1	1.175	96.77%
124	300	1/880	189/220	1	1.2	97.71%
119	300	1/905	786/905	1	1.2	97.59%
124	300	0.000852273	831/880	1.05	1.2	97.88%
124	275	0.000993377	0.926821192	1.05	1.2	97.47%
124	250	1/630	253/315	1	1.2	98.24%
150	290	1/700	11/14	1	1.2	97.06%

Table 2. Linear alpha function with fixed alpha range

Flpitch and fhpitch are the lower and upper frequency limits over which the linear value of alpha is applied. These values allow for the pitch distribution to have some influence over the weighting of the value of alpha. Three representative alpha functions from Table 2 are graphed in Figure 2. Each is constant up to flpitch, linear between flpitch and fhpitch, and then constant from fhpitch onwards.

Next, we select different fixed frequency ranges over which to test several different alpha value ranges. Numerous combinations of fixed frequency ranges were incrementally tested, but only the most representative range is shown in Table 3. The pitch frequency ranges are 150-290 Hz, 124-300 Hz, and 190-300 Hz. For each of these three pitch ranges, we show five different alpha ranges: 1.05-1.2, 1-1.19, 1-1.21, 1.05-1.19, and 1.05-1.21.

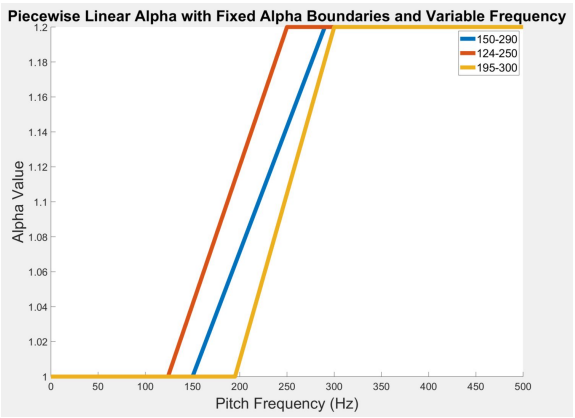

Fig. 2. Piecewise linear alpha values with fixed boundaries

Linear Alpha Clean Female with Variable Alpha						
Flpitch	Fhpitch	alphaL	alphaH	slope	intercept	Clean Female
150	290	1.05	1.2	1/933	249/280	97.88%
150	290	1	1.19	1/737	223/280	97.00%
150	290	1	1.21	1/667	31/40	96.94%
150	290	1.05	1.19	0.001	9/10	97.82%
150	290	1.05	1.21	1/875	123/140	97.65%
124	300	1.05	1.2	0.00085	831/880	98.18%
124	300	1	1.19	1/926	110/127	97.71%
124	300	1	1.21	1/838	812/953	97.88%
124	300	1.05	1.19	0.0008	802/843	98.00%
124	300	1.05	1.21	0.00091	777/829	97.41%
190	300	1.05	1.2	1/733	87/110	96.77%
190	300	1	1.19	1/579	565/841	92.48%
190	300	1	1.21	1/524	318/499	92.77%
190	300	1.05	1.19	1/786	316/391	96.59%
190	300	1.05	1.21	1/688	270/349	96.94%

Table 3. Linear alpha function with variable alpha range

The motivation behind the 150 Hz to 290 Hz frequency range is as follows. In Figure 1, the male pitch PDF has a slight peak in the 'tail' of the curve in the 150 - 200 Hz range. This extra peak is likely an artifact of small sample size and would no longer be present in a larger data set. We assume that the male probability distribution in this range would have continued with the same shape as prior. If so, the intersection of the male and female pitch PDFs would occur at ~150 Hz. On the high end, the maximum detected pitch in the clean female testing data set rounds to ~290 Hz. The mean pitch of the male training data set is ~124 Hz.
As seen in Table 3, although the best results occur with a range of 124-300 Hz with an alpha varied linearly between 1.05 to 1.2, we will use an alpha range of 1.0 to 1.2. This is because male pitch is more likely than female pitch to occur in the range between 124 to 150 Hz and we only warp female pitch.

To influence the alpha value dynamically, we interpolate the female pitch PDF and use the interpolation to determine a piecewise linear version of alpha. We ultimately select a piecewise cubic spline interpolation of the PDF shown in Figure 1. We calculate alpha as shown in Table 4. These are the formulas used to generate the alpha function. For each testing speech utterance, the median pitch is calculated. Based on where this pitch falls in the piecewise function of alpha in Table 4, alpha is either 1.0, determined by the equation, or 1.2.

The graph in Figure 3 shows how the alpha piecewise function in Table 4 modifies the slope of the baseline version of alpha. Although we tried higher order polynomial fits rather than piecewise cubic splines, instabilities and unnecessary complexities arise.

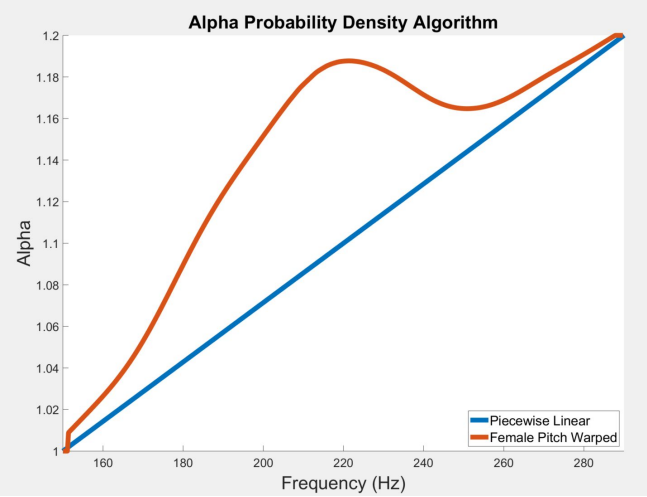

Fig. 3. Graph of dynamic alpha versus linear alpha

We have been able to significantly improve the clean female gender mismatch by applying VTLN with an alpha value that is weighted by the probability density distribution of clean female testing pitch. The final result is 97.82% accuracy with this method of VTLN. That is an improvement of 15.1% over the baseline of 82.72%. Most importantly, we demonstrate that the female pitch PDF can be used to effectively influence the alpha value for VTLN.

FORMULA	DESCRIPTION
$pp = pchip(Pitch, pdf\_FemPitch)$	*Piecewise cubic spline interpolation of female pitch pdf* (struct)
$pp.breaks = \begin{bmatrix} \\ \vdots \\ \end{bmatrix}$	*Pitch frequency vector*
$pp.coeffs = \begin{bmatrix} \cdots \\ \vdots \\ \cdots \end{bmatrix}$	*Cubic spline interpolation coefficients ; [a,b,c,d] for each frequency interval*
$fun = a(x\text{-}x_1)^3 + b(x\text{-}x_1)^2 + c(x\text{-}x_1) + d$	*Polynomial equation on interval* $[x_1 x_2]$
$pt = median\_pitch$	*Median pitch calculated for any given speaker*
$alpha\ baseline = \frac{1}{700}\ pt(p) + (\frac{11}{14})$	*Linear alpha warping function between 150 Hz and 290 Hz used as baseline alpha function*
$alpha = \begin{cases} 1 \\ \left(\frac{1}{700} + \frac{\text{abs(fun)}}{35}\right) \times pt(p) + \frac{11}{14} \\ 1.2 \end{cases}$	$pt < 150$ $150 \leq pt \leq 290$ $pt > 290$

Table 4. Dynamic alpha equation algorithm

3.2. Noise Reduction

In this section, we uncover two main noise reduction algorithms. The first algorithm is SCNR by Hendriks and the second is the Wiener-Scarlart Noise reduction algorithm [9].

3.2.1. Single Channel Noise Reduction (SCNR)

Speech enhancement refers to many methods that are designed to improve certain speech parameters, i.e. echo control bandwidth extension, packet loss and additive noise. This speech enhancement technique is based on Single-Channel additive noise reduction that works in the discrete Fourier transform (DFT) domain [1]. We will cover three implementations of Hendriks' Single-Channel noise reduction algorithm. First, we will investigate the types of speech DFT estimators. Second, we will cover noise Power Spectrum Density (PSD) estimation and finally, speech PSD estimation.

There are two different estimators that have been developed over the years: the complex-DFT (CDFT) and the magnitude DFT (MDFT). The CDFT is the DFT signal that composes the real and imaginary part, while MDFT is computing the magnitude of the DFT signal. Hendriks argued that since speech DFT histograms resemble more of a super-Gaussian curve with higher peaks and heavier tails. Thus, the MDFT estimator will be better for speech enhancement estimation (Figure 4). As a result, a linear MDFT with the minimum mean square estimation (MMSE) technique is applied in his algorithm to improve the *a priori* signal-to-noise ratio [1].

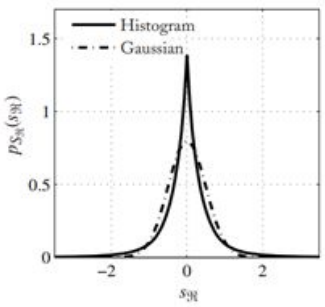

Fig. 4. Speech DFT Histogram

The earliest method of noise PSD estimation is the voice activity detection (VAD). However, it has been found that VAD determines speech presence by comparing the log-energy of the noise to adjust a threshold. This VAD-based approach cannot follow quick changes in the noise level, as shown in Figure 5.

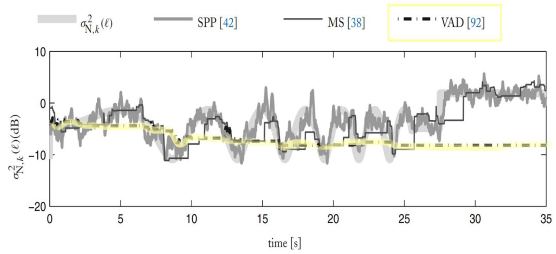

Fig. 5. Estimation of noise PSD

As a result, speech presence probability (SPP) is used in the SCNR code to determine the likelihood of speech presence and absence. It can then better estimate the noise PSD. A simple absence of speech between syllables or certain words. These moments enable us to estimate the noisy signal. To detect these non-speech parts, the Wiener-Scarlart filter for our project implements VADs. Similar to speech PSD estimation in the SCNR algorithm, the Wiener-Scarlart filter will also incorporate the decision-direct method [12].

3.2.3. Comparison of SCNR to Wiener Noise reduction

Aside from Speech PSD estimation, SCNR and Wiener Noise reduction have different noise reduction methods, see Table 6. Therefore, these noise reduction algorithms will tackle distinct aspects of noise differently.

	DFT estimation	Noise PSD estimation	Speech PSD estimation
Single-Channel Noise Reduction	Magnitude-DFT (MDFT)	Speech Presence Probability (SPP)	Decision-Direct (DD)
Wiener-Scarlart filter	Complex-DFT (CDFT)	Voice activation detection (VAD)	Decision-Direct (DD)

Table 6. Comparison of noise reduction methods

3.2.4. Results of Noise Reduction

speech PSD estimator is the maximum likelihood (ML) estimator. Since ML does not contain any smoothing, it causes spectral peaks that leads to musical like noise. Thus, to get a better trade-off, Ephraim and Malah proposed a decision-directed estimator [10]. By applying decision-directed estimators that allow better smoothing, the critical properties of speech are better preserved.

3.2.2. Wiener-Scarlart Noise Reduction

Hendriks proposed that since speech DFT is a super Gaussian distribution, the magnitude-DFT will be a better fit. However, if the speech and noise DFT resembles a regular Gaussian distribution, then the Wiener filter is optimal amongst all estimators [1]. We will divide the Wiener noise reduction explanation into three sections to cover the basic understanding of the Wiener-Scarlart speech enhancement: DFT estimator, Noise PSD estimation, and Speech PSD estimation.

Originally, the role of phase was thought to be less important. Krawczyk and Gerkmann proposed an algorithm able to estimate clean speech phase in a noisy signal [11]. As a result, the Wiener filter utilizes complex-DFT with a MMSE. Early methods for noise PSD estimation uses the fact that there is an

Feature	MFCC	MFCC	MFCC	MFCC	PNCC	PNCC	PNCC	MFCC
Method	Baseline	Spectral Subtraction	SCNR	Wiener NR	Baseline	SCNR	WienerNR	WienerNR then SCNR
Clean Male	99.4	99.32	99.13	99.13	98.69	98.51	98.32	98.69
Clean Female	82.72	81.25	83.42	80.66	64.08	63.84	61.9	79.95
10dB Male	82.45	77.29	84.63	87.8	91.16	90.42	88.43	86.75
10dB Female	36.51	37.33	40.62	47.62	43.86	43.74	43.27	47.68
5dB Male	45.43	43.44	57.69	68.01	77.29	79.28	72.06	65.53
5dbB Female	12.35	17.28	24.81	30.69	29.1	33.63	30.92	30.1
Average	59.81	59.32	65.05	68.99	67.36	68.24	65.80	68.12

Table 5. Final results

In this section, we explore the results from the above implementation for noise reduction. We evaluate different combinations of noise reduction methods on noisy speech (Figure 6A). From Table 5 results, if a Wiener-Scarlart filter is used before SCNR, the results are worse than just using a Wiener-Scarlart filter alone. This may be due to the Wiener filter overestimating the SNR. With additional noise algorithms in addition to PNCC, the results slightly degrade. Since PNCC does not perform well in gender mismatch, the overall results are adversely affected. However, if gender mismatch is not taken into account, PNCC has the best results in terms of noise reduction.

3.3. Overall Results

Overall, MFCC with SCNR followed by Wiener-Scarlart filter yields the best results with an average score of 81.19% (Table 5). Clean female drops from a maximum of 97.82% for dynamic VTLN (D-VTLN in Table 5 and is the dynamic alpha determining method we developed) to 94.12%. Also, neither 10dB male, 5dB male or clean male achieve their maximums

with MFCC with SCNR, Wiener-Scarlart filter and D-VTLN. However, the improvements obtained for 10dB and 5dB female are significantly better than for any other condition. Additionally, the differences between male and female for 5dB male and female as well as 10dB male and female are smallest for the combination of MFFC + SCNR + Wiener + D-VTLN.

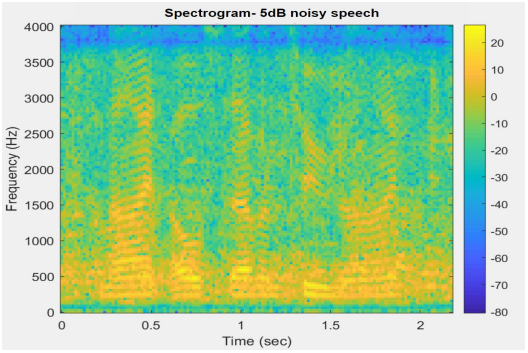

Fig. 6A. Spectrogram 5dB noise of words "90729"

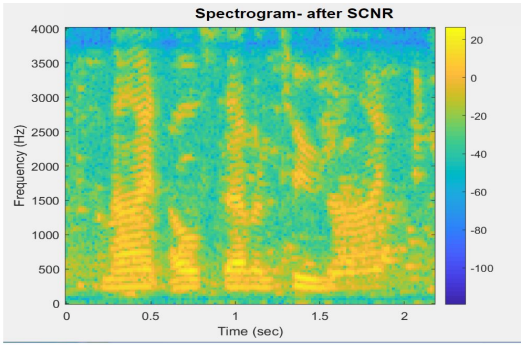

Fig. 6B. Spectrogram after applying SCNR

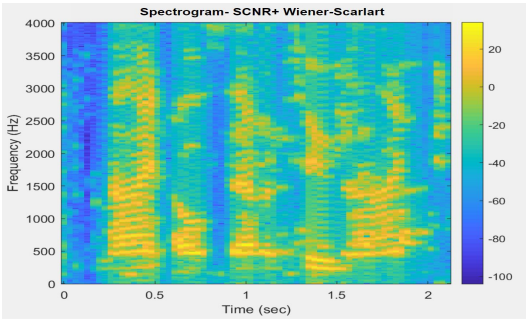

Fig. 6C. Applying SCNR and Wiener-Scarlart filter

4. SUMMARY AND DISCUSSION

We set out with two objectives: uncover a new way to determine the warping factor alpha for standard VTLN implementations with MFCC feature extraction and find cutting edge or lesser known methods for reducing noise during feature extraction to improve a HTK HMM machine learning isolated word recognizer for connected digits. We successfully do both. For VTLN, we create an algorithm for dynamically determining alpha based on the probability density distribution of female pitch. In this report, the female pitch probability distribution was generated from the clean female testing data. However, for a more accurate representation in future work, the probability distribution of female pitch data would be from a much larger sample size. The basis for the determination of alpha relies on the following: 1) setting boundaries for the dynamic scaling of alpha 2) determining the pitch PDF of the clean female testing data 3) using the PDF to influence the slope of alpha. After exploring a range of methods for noise reduction, we select a SCNR filter combined with a Wiener-Scarlart filter to address the challenges of noise reduction. The total combination of our system improves performance in all six testing categories.

There are many opportunities for continued improvement on our work. A primary issue is the length of time our software takes to process the feature extraction algorithms. Both the gender and noise reduction sections of the code could be rewritten to reduce processing time. Although the gender mismatch resolution addresses a new approach that uses female pitch information to select the alpha value for VTLN, the algorithm needs further refinement and exploration. The filters complement each other to effectively remove noise but implementing both the SCNR and Wiener-Scarlart filters adds complexity that may be undesirable. In summary, we achieve our goal of finding a new way to dynamically determine the alpha value of a standard VTLN as well as find powerful and lesser known noise reduction filter that together improve the performance of an isolated word recognizer for connected digits operating under the constraints of noise and gender mismatch.

5. REFERENCES

[1] R.C. Hendriks, T. Gerkmann, and J. Jensen, "DFT-Domain Based Single-Microphone Noise Reduction for Speech Enhancement," Morgan & Claypool, Delft University of Technology, 2013.

[2] A.P. Simpson, "Phonetic differences between male and female speech," *Language and Linguistics Compass,* Vol. 3, pp.621-640, March 2009.

[3] Y. Takefuta, "A statistical analysis of melody curves in the intonation of American English," *Proceedings of the 7th International Congress of Phonetic Sciences,* pp.1035-1039, 1972.

[4] D. Stadniczuk, G. Bauckmann, and D. Suendermann-Oeft, "An Open-Source Octave Toolbox for VTLN-Based Voice Conversion," *International Conference of the German Society for Computational Linguistics and Language Technology,* 2013.

[5] S. Panchapagesan, A. Alwan, "Frequency Warping for VTLN and Speaker Adaptation by Linear Transformation of Standard MFCC," *Computer Speech & Language,* Vol. 23, Issue 1, pp. 42-64, January 2009.

[6] S. Young, G. Evermann, M. Gales, et al, "The HTK Book," Cambridge University Engineering Department, December 2015.

[7] N. Upadhyay and A. Karmakar, “Single channel speech enhancement utilizing iterative processing of multi-band spectral subtraction algorithm,” *Power, Control and Embedded Systems (ICPCES) 2nd International Conference*, Dec 2012.

[8] L. N. Tan and A. Alwan, “Multi-Band Summary Correlogram-based Pitch Detection for Noisy Speech,” *Speech Communication,* Vol. 55, Issues 7-8, pp. 841-856, September 2013.

[9] R. Hendriks, “Algorithm for Noise reduction for speech enhancement,” https://www.mathworks.com/matlabcentral/fileexchange/46171-algorithm-for-noise-reduction-for-speech-enhancement?s_tid=prof_contriblnk, MathWorks, 2014.

[10] Y. Ephraim, “Statistical-model-based speech enhancement systems,” *Proc. IEEE*, Vol. 80, No. 10, pp. 1526–1555, Oct. 1992.

[11] M. Krawczyk and T. Gerkmann, “STFT phase improvement for single channel speech enhancement,” *Int. Workshop Acoustic Echo, Noise Control*, Aachen, Germany, Sep. 2012

[12] P. Scarlart and J. Vieira Filho, “Speech enhancement based on Priori signal to noise estimation”, *IEEE*, 1996.