Marcin PŁONKOWSKI[1], Pavel URBANOVICH[1,2]

The John Paul II Catholic University of Lublin, Poland (1), Belarusian State Technological University, Minsk (2)

doi:10.15199/48.2016.08.21

The use of pitch in Large-Vocabulary Continuous Speech Recognition System

Abstract. *In this article the authors normalize the speech signal based on the publicly available AN4 database. The authors added to the algorithm of calculating the MFCC coefficients, the normalization procedure, that uses pitch of the voice. As demonstrated by empirical tests authors were able to improve speech recognition accuracy rate of about 20%.*

Streszczenie. *W niniejszym artykule autorzy normalizują sygnał mowy wykorzystując publicznie dostępną bazę danych AN4. Autorzy dodali do algorytmu obliczania współczynników MFCC, procedurę normalizacji, wykorzystującą wysokość tonu głosu. Jak wynika z przeprowadzonych testów, autorzy uzyskali poprawę dokładności rozpoznawania mowy o około 20% (**Wykorzystanie wysokości tonu głosu w systemach rozpoznawania mowy ciągłej z dużą ilością słów**).*

Keywords: speech recognition, CMU Sphinx, pitch, speech normalization.
Słowa kluczowe: rozpoznawanie mowy, CMU Sphinx, wysokość głosu, normalizacja sygnału mowy.

Introduction

The fundamental frequency (*F0*) plays a very important role in generating speech signal [1]. It has long been known that, formant frequencies generated by women and men are different from each other. Women have higher formant frequencies than men [2], which may be explained by the longer vocal tracts of men [3]. The important source of the inter-speaker variations is the vocal tract length (VTL). Therefore, vocal tract length normalization (VTLN) technique was described in many publications [4 – 6].

However, information about VTL is rarely used by speech recognition systems. Most often they use MFCC coefficients for each frame calculated in the same way. This means that continuous speech recognition systems do not take into account the changes in the spectrum. These changes appear in the utterance of the same phoneme spoken by various speakers.

One of the most commonly used technique for the normalization is bilinear transform (BLT) [7-9]. It was presented in [10] using BTL in CMU Shinx (speech recognition system). However, this technique introduces additional overhead when calculating the cepstral coefficients.

In this article, we will show a simple and very effective method of normalization of the speech signal which predisposes it to be used for mobile devices.

Pitch detection

In this article we will carry out normalization based on the sex of the speakers. As mentioned earlier (due to differences in the length of the vocal tract), women have higher pitch and formant frequencies than men. Typical values obtained for *F0* are 120 Hz for men and 210 Hz for women [11]. In general, both voice pitch and formant frequencies were lower in men than in women [12]. Thus, a key issue in the normalization process will be the estimation of the pitch of the speaker.

To detect a sound's pitch, the algorithm performs an acoustic periodicity detection on the basis of an accurate autocorrelation method, as described in [13]. This method is very accurate, noise-resistant, and robust.

For a time signal *x(t)*, the autocorrelation *r(z)* as a function of the lag *z* is defined as:

$$r(z) = \int x(t)x(t+z)dt \,. \tag{1}$$

This function has a global maximum for $z = 0$. If there are also global maxima outside 0, the signal is called periodic and there exists a lag *z*, called the period. The pitch of the signal *x(t)* is defined as $F0 = 1/Z$. This method is used in Praat which is a free scientific computer software package for the analysis of speech in phonetics.

To detect a sound's pitch we use autocorrelation, comparing each window with itself. An autocorrelation plot shows the degree to which the compared curves are related on the Y-axis, and the time lag for each comparison on the X-axis. If the curve is periodic, then there should be a peak on the autocorrelation curve when the lag is equal to the curve's period. The autocorrelation is highest, when a time lag is equal to 0. Thus, we need to look for peaks that are greater than 0 (see Fig. 1). The autocorrelation curve shows a false peak (solid vertical line) before the time lag, that is, the sound's actual fundamental frequency (dashed vertical line) [14].

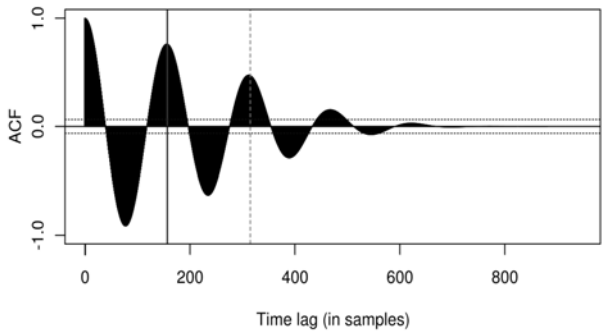

Fig.1. Normalized autocorrelation of the filtered sound [14]

In order to correct for this, we divide the signal by the normalized autocorrelation curve of the windowing function. And by finding the maximum at a time lag > 0 (see Fig. 2), we can compute the pitch of the signal converting from samples to Hz [14].

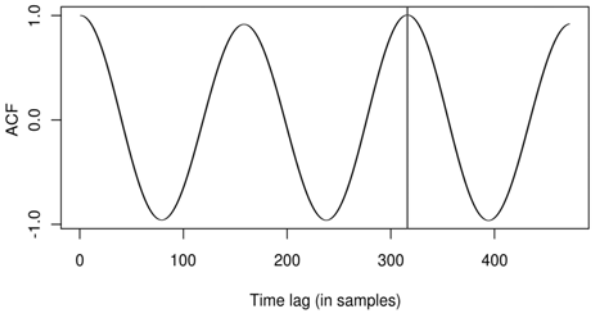

Fig.2. Estimated autocorrelation of the original signal [14]

Table 1. Average formant frequencies in Hz for men, women and children for a selection of vowels [2].

Vowel in	Men			Women			Children		
	F1	F2	F3	F1	F2	F3	F1	F2	F3
beat	270	2300	3000	300	2800	3300	370	3200	3700
bit	400	2000	2550	430	2500	3100	530	2750	3600
bet	530	1850	2500	600	2350	3000	700	2600	3550
bat	660	1700	2400	860	2050	2850	1000	2300	3300
part	730	1100	2450	850	1200	2800	1030	1350	3200
pot	570	850	2400	590	900	2700	680	1050	3200
boot	440	1000	2250	470	1150	2700	560	1400	3300
book	300	850	2250	370	950	2650	430	1150	3250
but	640	1200	2400	760	1400	2800	850	1600	3350
pert	490	1350	1700	500	1650	1950	560	1650	2150

We can calculate fundamental frequency *F0,* using the formula:

$$F0 = \frac{1}{lag_{\max} / f_s}, \tag{2}$$

where: f_s is sampling rate.

Formants shift

A consequence of higher fundamental frequency in female voice (compared to male voice) is the formant frequency offset (see Tab. 1). These shifts are not linear. Changes within the first formant are the smallest. They vary from 10 Hz to 200 Hz. The mean change is 70 Hz. Changes within the second formant vary from 50 Hz to 500 Hz. The mean change is 275 Hz. Changes within the third formant vary from 250 Hz to 550 Hz. The mean change is 395 Hz.

Children's voices are characterized by even greater shifts due to the higher fundamental frequency. However, in our tests, we consider only the voices of adult speakers.

Frequency normalization

In our analysis we use CMU Sphinx III. This is one of the most popular speech recognition systems. It works very well in continuous speech recognition tasks with a lot of words, regardless of the speaker. However, to achieve satisfactory results, the system must be trained on the appropriate set of utterances with the reference transcription.

The whole process of speech recognition by decoder starts with acquisition of utterance. Then, the extraction process is performed of the most desirable features (from the point of view of speech recognition system). Decoder analyzes these features using acoustic model, language model and vocabulary. Block diagram is shown in Fig.3.

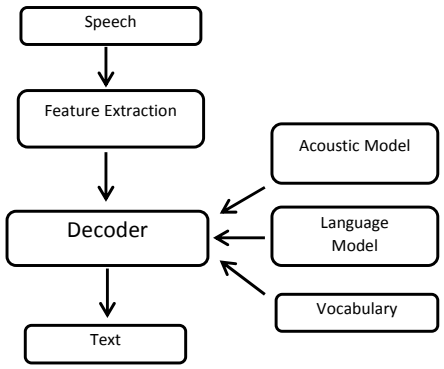

Fig.3. Block diagram of speech recognition system

CMU Sphinx-III is a system that uses statistical methods. Namely, this system is based on a hidden Markov model (HMM). It is now the dominant solution for the most recently designed speech recognition systems. If we have a good learning set (of appropriate size and of appropriate quality) the system gives very good results (word error rate is approximately 15%).

To obtain very good results training set size should take into account the following recommendations:

- 1 hour of recording for command and control for single speaker
- 5 hours of recordings of 200 speakers for command and control for many speakers
- 10 hours of recordings for a single speaker dictation
- 50 hours of recordings of 200 speakers for many speakers dictation

We briefly describe the signal processing front end of the SPHINX III speech recognition system. The front end transforms a speech waveform into a set of features to be used for recognition, specifically, mel-frequency cepstral coefficients (MFCC).

The front end processing performed by the Sphinx-III:

- pre-emphasis
- windowing (Hamming window)
- power spectrum
- mel spectrum
- mel cepstrum.

Sphinx III uses the following features:

- Sample rate: 16000 Hz
- FFT Size: 512
- Frame Size: 410
- Window size: 25ms
- Window shift: 10ms
- Pre-emphasis coefficient: 0.97

Sphinx III uses the following MFCC features:

- 12 MFCC (mel frequency cepstral coefficients)
- 1 energy feature
- 12 delta MFCC features
- 12 double-delta MFCC features
- 1 delta energy feature
- 1 double-delta energy feature
- Total 39-dimensional features

The Sphinx III uses MFCC coefficients for each frame calculated in the same way. We will process normalization procedure after the calculation of the power spectrum.

As a base spectrum we assume male voice spectrum. And when we detect female voice we make a shift in the frequency domain. As mentioned earlier, typical values obtained for *F0* are 120 Hz for men and 210 Hz for women.

Therefore, if the frequency *F0* is greater than 165 Hz we will consider the voice as a female, and if *F0* is less than 165 Hz as a male voice.

If we qualify utterance as a male voice, the power spectrum will remain unchanged. But, If we qualify utterance as a female voice, the power spectrum will be shifted in the direction of lower frequencies.

Tests and results

In this article we normalize the speech signal based on the publicly available AN4 database. The database has 948 training and 130 test utterances. All data are sampled at 16 kHz, 16-bit linear sampling. All recordings were made with a close talking microphone.

The directory with training data has 74 sub-directories, one for each speaker. 21 of them are female, 53 are male. The total number of utterances is 948, and the average duration is about 3 seconds, totaling a little less than 50 minutes of speech. The directory with test data has 10 sub-directories, one for each speaker. 3 of them are female, 7 are male. The total number of utterances is 130, totaling around 6 minutes of speech [15].

The file structure for the database is:

- etc
  - an4.dic - Phonetic dictionary
  - an4.phone - Phoneset file
  - an4.lm.DMP - Language model
  - an4.filler - List of fillers
  - an4_train.fileids - List of files for training
  - an4_train.transcription - Transcription for training
  - an4_test.fileids - List of files for testing
  - an4_test.transcription - Transcription for testing
- wav
  - speaker_1
    - file_1.wav - Recording of speech utterance
  - speaker_2
    - file_2.wav

We will carry out two tests. In the first test we will try to find the optimal shift in a group of women's voices, ensuring the lowest rate mistakenly recognized words (WER – word error rate).

We estimate the accuracy of using number of incorrectly recognized words WER (word error rate), which is defined as:

$$WER = \frac{S + I + D}{N}, \tag{3}$$

where: S is the number of substitutions, I is the number of insertions, D is the number of deletions, N is the number of words in the reference.

The word error rate (WER) is the most common way to evaluate speech recognizers. The word error rate is defined as the sum of these errors divided by the number of reference words. It is worth noting that according to the formula (3) WER value may be greater than 100%.

When reporting the performance of a speech recognition system, sometimes word accuracy (WAcc) is used instead:

$$WAcc = 1 - WER. \tag{4}$$

The sampling rate was set to 16,000 Hz. Thus, each frame contains 410 samples. The frame shift is 160 samples. The FFT size parameter will be set to 512.

For a sampling frequency of fs = 16,000 Hz and a signal of length N = 512 points (FFT size), the frequency resolution is fs/N = 16,000/512 = 31.25 Hz. Thus magnitudes will show up in the spectrum at 0 Hz, 31.25 Hz, 62.50 Hz, 93.75. Hz… up to 8000 Hz and at only these frequencies, so that all other frequencies are undefined. Thus only the multiple of 31.25 Hz can be used as the shift values.

The error for the baseline system is equal to 15.27 (WER). In the first analysis we will make the same shift in the frequency domain. We analyze the amount of shift in the range of 31.25 Hz to 650 Hz in Increments of 31.25 Hz. In table 2 we show errors depending on the shift amount. As seen in Table 2, we achieve the best results when shifting about 187.5 Hz. The best result is 12.55 (WER).

When we compare this result with the result of the baseline system (15.27), we can notice that we achieve an increase speech recognition accuracy by about 17.81 %.

The data in Table 2 are also shown on the graph (see Fig. 4). It represents the values WER depending on shifts in the frequency domain.

However, as seen in Table 1 shifts in the frequency domain are not linear.

For example, $F1$ formant frequency for the phoneme /i:/ in the word “beat” is about 270 Hz for male voices and 300 Hz for female voices. Thus, the difference is 30 Hz. $F2$ formant frequency for the same phoneme /i:/ in the word “beat” is about 2300 Hz for male voices and 2800 Hz for female voices. Thus, the difference is 500 Hz. $F3$ formant frequency for the same phoneme /i:/ in the word “beat” is about 3000 Hz for male voices and 3300 Hz for female voices. Thus, the difference is 500 Hz.

Therefore, we see that a more optimal way is to create a number of ranges with different shifts in the frequency domain.

Based on analysis of data from Table 1, we propose the following ranges:

- 0 Hz – 593.75 Hz, shifted of 62.5 Hz,
- 625 Hz – 1250 Hz, shifted of 187.5 Hz,
- 1281.25 Hz – 1750 Hz, shifted of 281.25 Hz,
- 1781.25 Hz – 8000 Hz, shifted of 500 Hz.

In our second test, we found that the word error rate was 11.51%. When we compare this result with the result of the baseline system (15.27%), we can notice that we achieve an increase in speech recognition accuracy by about 24.62%.

As seen in Table 3, we compared our results with the baseline system. The data in Table 3 are also shown on the graph (see Fig. 5). The horizontal line is the result of the baseline system – 15.27% WER.

Table 2. Word error rate (WER) depending on the shift amount

The shift amount (in Hz)	31.25	62.5	93.75	125	156.25	187.5	218.75	250	281.25	312.5
WER (%)	13.84	13.58	14.49	13.45	13.07	12.55	14.62	15.14	15.78	15.65
The shift amount (in Hz)	343.75	375	406.25	437.5	468.75	500	531.25	562.5	593.75	625
WER (%)	16.95	16.95	15.14	15.01	15.91	18.76	16.82	16.95	15.78	17.85

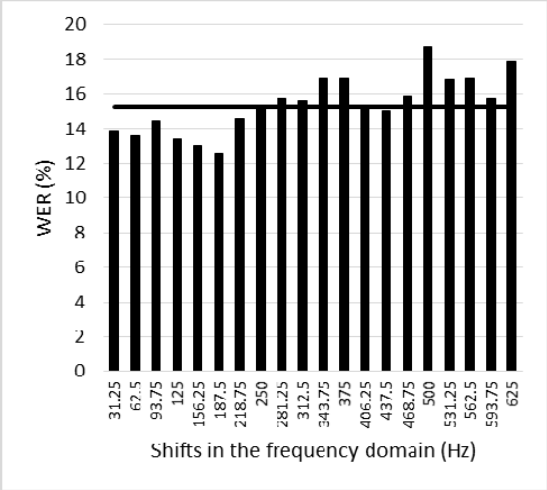

Fig. 4. The WER values depending on shifts in the frequency domain (the horizontal line is the result of the baseline system – 15.27% WER).

Table 3. Comparison of our results with the baseline system

	Baseline system	Test 1.	Test 2.
WER (%)	15.27	12.55	11.51
Improvement in %	–	17.81	24.62

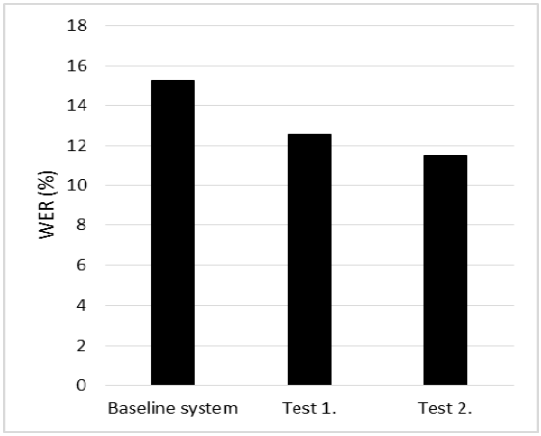

Fig. 5. Comparing of the two tests with respect to the baseline system

Conclusions

In this article we analyzed the difference in the speech signal between men and women. Due to the higher frequency of formants in the female voices, they made shifts in the frequency domain. In the first test, the authors made a fixed shift across the frequency range. It turned out that the best result (12.55% WER) they had achieved for shifting 187.5Hz. It gives the increase in accuracy of 17.81%, with regard to the baseline system.

Due, to the fact that the shifts in the frequency domain are not linear, we create a number of ranges with different shifts in the frequency domain. It turned out that in the second test we obtained improvements in speech recognition accuracy equal to 11.51%. It gives the increase in accuracy of 24.62% with regard to the baseline system.

Additional advantage of this normalization is the efficiency because it does not introduce any noticeable overhead. Thus, it can be successfully used e.g. in mobile devices.

In our future work we will analyze this normalization procedure regarding the Polish language in which some elements are different from English [16, 17]. Hence they require separate analysis.

Authors: *dr Marcin Płonkowski, Katolicki Uniwersytet Lubelski Jana Pawła II, Instytut Matematyki, Katedra Systemów Operacyjnych i Sieciowych, ul. Konstantynów 1H, 20-708 Lublin, E-mail: marcin.plonkowski@kul.lublin.pl; prof. dr hab. Paweł Urbanowicz, Katolicki Uniwersytet Lubelski Jana Pawła II, Instytut Matematyki, Katedra Systemów Operacyjnych i Sieciowych, ul. Konstantynów 1H, 20-708 Lublin, E-mail: pav.urb@yandex.by*

REFERENCES

[1] Benesty J., Sondhi M.M., Huang Y., Springer Handbook of Speech Processing, Springer, Berlin, 2008
[2] Peterson G.E., Barney H.L., Control methods used in a study of the vowels, Journal of the Acoustical Society of America, 24 (1952), 175–184
[3] Fitch W.T., Giedd J., Morphology and development of the human vocal tract: a study using magnetic resonance imaging, Journal of the Acoustical Society of America, 106 (1999), n.3, 1511-1522
[4] Zhan, P., Waibel, A., Vocal Tract Length Normalization for Large Vocabulary Continuous Speech Recognition, CMU-CS-97-148, Carnegie Mellon University, Pittsburgh, PA. May. 1997
[5] Kamm T., Andreou G., Cohen J., Vocal Tract Normalization in Speech Recognition: Compensating for Systematic Speaker Variability, Proc. of the 15th Annual Speech Research Symposium, pp. 161-167, CLSP, Johns Hopkins University, Baltimore, MD, June 1995
[6] Tuerk C., Robinson T., A new frequency shift function for reducing inter-speaker variance. In: Proc. Eurospeech 1993, 351-354
[7] Oppenheim A.V., Johnson D.H., Discrete representation of signals. Proc. of the IEEE, 60 (1972), n.6, 681-691
[8] Acero, A., Acoustical and Environmental Robustness in Automatic Speech Recognition. PhD thesis, Carnegie Mellon University, Pittsburgh, Pennsylvania, 1990
[9] McDonough, John W., Speaker Compensation with All-Pass Transforms. PhD thesis, Johns Hopkins University, 2000
[10] Waibel A., Lee K.F., Readings in Speech Recognition, Morgan Kaufmann, 1990
[11] Traunmüller, H., Eriksson, A., The frequency range of the voice fundamental in the speech of male and female adults. http://www.ling.su.se/staff/hartmut/aktupub.htm
[12] Johnson, K. Speaker Normalization in speech perception. In Pisoni, D.B. & Remez, R. (eds) The Handbook of Speech Perception. Oxford: Blackwell Publishers, (2005), 363-389
[13] Boersma P., Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound, Proceedings of the Institute of Phonetic Sciences University of Amsterdam, 17 (1993), 97-110
[14] A brief explanation of Praat's pitch detection algorithm, http://www.ucl.ac.uk/~ucjt465/tutorials/praatpitch.html
[15] The CMU Audio Databases, AN4 database, http://www.speech.cs.cmu.edu/databases/an4/
[16] Płonkowski M., Urbanowicz P., Tuning a CMU Sphinx-III Speech Recognition System for Polish Language, Przegląd Elektrotechniczny (2014), n.4, 181-184
[17] Płonkowski M., Using bands of frequencies for vowel recognition for Polish language, International Journal of Speech Technology, 18 (2014), n.2, 187-193

.