The Use of Pitch in Large-Vocabulary Continuous Speech Recognition System

Paweł Urbanowicz, Marcin Płonkowski
The John Paul II Catholic University of Lublin, Lublin, Poland,
E-mail: marcin.plonkowski@kul.pl

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].

However, information about F0 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 popular speech recognition systems as CMU Sphinx, does not allow the use of the information about the pitch. However, thanks to the openness of the source code, we can make this modification and try to normalize the signal.

It has already been published several attempts to analyze height of formant frequencies depending on gender [4] and proposed several algorithms for the normalization of the voice signal due to the pitch [5]. These algorithms usually refer to vowels or only the voiced portions of the speech signal. There are not many publications showing the use of normalized frames in continuous speech recognition tasks.

In this article the authors have to try to normalize the speech signal based on the publicly available AN4 database [6]. The authors added to the algorithm of calculating the MFCC coefficients, the normalization procedure due to the pitch. As demonstrated by empirical tests authors were able to improve speech recognition accuracy rate of about 20%.

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, p.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, vol. 106, n.3, p.1511-1522.
[4] Chládková K., Boersma P., Podlipský V. J.: *On-line Formant Shifting as a Function of F0*, Proceedings of Interspeech 2009 Brighton, p. 464-467.
[5] Fujimoto K., Hamada N., Kasprzak W.: *Estimation and tracking of fundamental, 2nd and 3d harmonic frequencies for spectrogram normalization in speech recognition*, Bulletin of the Polish Academy of Sciences Technical Sciences, vol. 60, n.1, 2012, p. 71-81.
[6] The CMU Audio Databases, AN4 database, http://www.speech.cs.cmu.edu/databases/an4/.