Krajowa Konferencja Zastosowań Matematyki w Biologii i Medycynie

Łochów, 23rd–27th September 2014

SPEECH NORMALIZATION TO AVERAGE SPEAKER

Mariusz Mąsior and Mariusz Ziółko

AGH University of Science and Technology,
Faculty of Computer Science, Electronics and Telecommunications
al. Mickiewicza 30, 30-059 Kraków,
{masior, ziolko}@agh.edu.pl

ABSTRACT

The paper presents a new approach to speech normalization. It concerns, in spite of traditional Vocal Tract Normalization methodology, suppression of differences in speakers vocal tracts by unification of amplitude spectra (not the changes in the distribution of signal energy over spectra). The technique for generation an artificial, universal speaker's features as a reference level for normalization is introduced, as well as procedures for normalization itself. The algorithm verification (conducted on Polish speech recordings) and evaluation results are presented.

INTRODUCTION

A huge development of voice technology and expand of opportunities for their applications has led to a number of ideas for improving various elements of speech technology. One of such elements is an attempt to normalize speech in order to eliminate the effects associated with a variety of speakers.

The variations in acoustic speech signals for different speakers are caused by different sizes of vocal tracts, gender, different accents, dialects, speaking rates, style influenced by speaker's personality and current emotional state. Vocal Tract Normalisation (VTN) is a procedure (or set of procedures) which is typically applied in speaker independent automatic speech recognition. VTN improves recognition accuracy [1], which results in better efficiency of Automatic Speech Recognition (ASR) systems. In our case, the purpose of applying the normalization methods is a modification of multilingual speech recordings in order to obtain better data for comparative analysis of worldwide languages [2, 3].

Differences in vocal tract are manifested through differences in spectra, even when the speakers are generating a sound of the same phoneme. Therefore, the speech normalization is achieved by modifying the spectral characteristics and energy distribution across spectrum (all of it carried out in a layer of acoustic signal – before the parameterization). The idea of VTN has been considered by numerous researchers for fifteen years and has resulted in many insights [4–11].

As mentioned, the VTN involves an appropriate modification of the speech signal, especially in the frequency domain. The variation of vocal tracts lengths is considered as a main reason for the differences in the speech for different people. This effect can be neutralized by applying frequency warping, which consist of mapping speech spectrum $\hat{s}(k)$ into normalized spectrum $\hat{s}(\varphi(k))$, in accordance to arbitrary generated frequency warping function $\varphi(k)$ [9, 12].

SPEAKER NORMALIZATION TECHNIQUE

The presented approach is inspired by general methodology of noise and distortion cancellation, by adaptation to channel in which the signal is transmitted [13]. Such approach always consists of attempt to develop a model of a transmission channel (described, for example, by transmittance) and building an inverse signal transformation, which will eliminate the effects introduced by the channel.

Let us assume that we have recorded training speech samples $s^i \in \mathrm{Re}^N$ for all of $i = 1, 2, \ldots, I$ speakers. These recordings should be long enough (lasting at least 10–20 seconds) to correctly represent the frequency properties of analysed speaker. All speakers can be characterized by the complex spectra $\hat{s}^1(k), \hat{s}^2(k), \ldots, \hat{s}^I(k)$ obtained by applying the Fourier transform

$$\hat{s}^i(k) = \sum_{n=0}^{N} s^i(n)\, e^{-2\pi jkn/N}. \tag{1}$$

to speech samples $s^1(n), s^2(n), \ldots, s^I(n)$.

Average spectrum can be generated for all I speakers to represent the universal speaker spectral characteristic

$$\bar{\hat{s}}(k) = \frac{1}{I} \sum_{i=1}^{I} \left|\hat{s}^i(k)\right|. \tag{2}$$

The transfer function for each speaker can be created as a ratio of the amplitude spectra

$$d^i(k) = \begin{cases} \frac{\bar{\hat{s}}(k)}{|\hat{s}^i(k)|} & if \quad \left|\hat{s}^i(k)\right| \geq \varepsilon \\ 1 & if \quad \left|\hat{s}^i(k)\right| < \varepsilon \end{cases}. \tag{3}$$

Normalization of i-th speaker's voice recording can be simply computed in the frequency domain

$$\underline{\hat{s}}^i(k) = d^i(k)\, \hat{s}^i(k). \tag{4}$$

Finally, the normalized speech signal $\underline{s}^i(k)$ may be obtained from the inverse Fourier transform of normalized spectrum, i.e.

$$\underline{\hat{s}}^i(k) \text{ for } k = 0, \ldots, N \quad \xrightarrow{IFFT} \quad \underline{s}^i(n) \text{ for } n = 0, \ldots, N. \tag{5}$$

All vectors (signals and transfer functions) in Eqs.(2)–(4) need to have the same lengths. It can be simply achieved by zero padding in time domain, which corresponds to ideal interpolation for the spectral representation [14].

The proposed method does not eliminate the shift of frequencies caused by different speakers, but will change signal spectrum envelope, allowing the normalization of the effects imposed by different voice transmission conditions (in vocal tract and in environment of signal propagation).

EVALUATION

Recordings for testing were obtained from Polish speech Corpora [15]. This collection contains 45 sets (each consists of 365 recordings) for 37 distinct voices. Each of the speech recordings has time annotation for phoneme segmentation.

The presented experiment evaluation used only male speakers recordings (28 speakers) because of more balanced recordings conditions and greater number of male then female speakers. Comparing speakers of different genders may be affected by a significant difference in the fundamental frequency (which can be compensated by other methods). Data set contains of 114 recordings for every speaker (only full phrases recordings were long enough to give reliable results).

The concatenation of five recordings (a total of about 15 s) for every speaker (the same for each one) was used for speaker's spectrum determination. All the remaining 109 recordings (approximately 250 s for every speaker) were used to test the effectiveness of the algorithm.

It is difficult to find an effective and reliable way to test normalization. The most intuitive and reliably way would be to integrate it in speech recognition system and measure the changes in Word Error Rate according to parameters of normalization. Unfortunately, such calculations are not possible for a wide range of parameters due to the long computation time.

The algorithm optimization or examination large number of parameters demands different measure, that can accurately represent the possibilities of normalization. Such a measure can be developed through the comparison of distances between the vectors of parameters of the same phoneme realization by various speakers. Regardless of the methods of operation, speech processing always involves measuring distances between the parameters of phonemes and their models from the referenced data.

Speech parameterization involves the representation of its spectrum in a way that can effectively represent the most relevant information. Parameterization for evaluation on our signals (original and normalized) will be performed as transformations

$$s(n) \xrightarrow{P} p(l,k), \tag{6}$$

$$\underline{s}(n) \xrightarrow{P} \underline{p}(l,k), \tag{7}$$

where l is a time variable and k is a number of frequency band. This parameterization is achieved by applying wavelet packet transform in which at all stages of Discrete Wavelet Transform both the low-pass and high-pass bands are split [16]. It allows analyzing speech in frequency ranges corresponding to perceptual scale.

Each phoneme after parameterization is represented by a series of vectors. The series length is dependent on the phoneme duration. Therefore, in order to reduce the amount of calculations, a stationary model of phonemes was adopted. Assuming that the boundaries of phonemes (for each recording) are specified in the vector τ, the parameters of each phonemes take the form of

$$p_m(k) = \sum_{l=\tau(m)}^{\tau(m+1)} w^W(l) p(l,k), \quad W = \tau(m+1) - \tau(m) + 1. \tag{8}$$

Eq. (8) performs a weighted average, where the weight was taken as a Hamming window

$$w^W(l) = 0.54 - 0.46 \cos\left(\frac{2\pi l}{W-1}\right). \tag{9}$$

The distance for m-th phoneme realizations between speakers i and j can be computed as a Euclidean distance of parameters

$$D_m^{i,j} = \sqrt{\sum_k \left(p_m^i(k) - p_m^j(k)\right)^2}. \tag{10}$$

A similar distance measure

$$\underline{D_m^{i,j}} = \sqrt{\sum_k \left(\underline{p_m^i}(k) - \underline{p_m^j}(k)\right)^2} \tag{11}$$

can be computed for signals after normalization, obtained from Eq. (7).

Only phonemes (derived from various speakers) with an identical context (i.e. an identical position in the recording) were used to compare distances, for additional limitation in the comparisons number and for possibility of different phonemes representations (stationary or temporal models).

An indicator of the normalization quality can be defined as a mean, relative reduction of referenced phonemes distances between a pair of speakers (i and j)

$$R^{i,j} = 1 - \frac{1}{M} \sum_{m} \frac{D_m^{i,j}}{\overline{\overline{D_m^{i,j}}}}. \tag{12}$$

Finally, the averaging of index $R^{i,j}$ for referenced speaker ($i-$th) will be considered as a normalization quality index

$$\bar{R}^i = \frac{1}{I-1} \sum_{j=1, j\neq i}^{I} R^{i,j}. \tag{13}$$

RESULTS

The average spectrum $\bar{\bar{s}}(k)$ (universal speaker characteristic) was calculated in accordance with Eq. (2) and it is shown in Fig. 1. The spectrum is typical for the speech signal — average position of formants can be easily seen (maxima of speech signal spectrum envelope). A chart of the

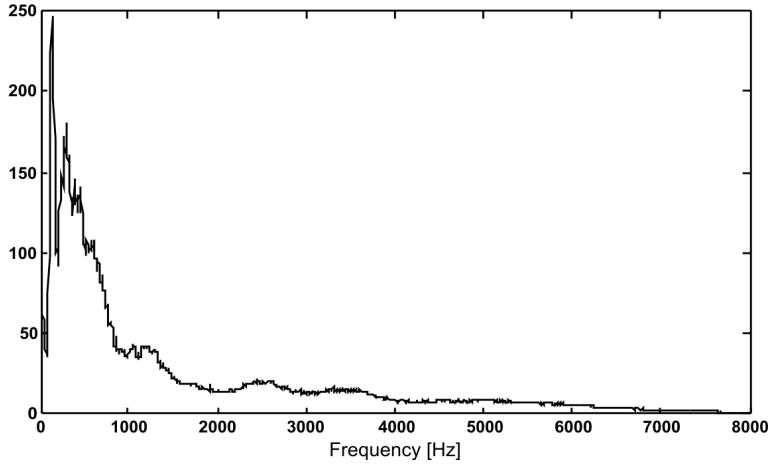

Figure 1. Amplitude-frequency characteristics of averaged signal (universal speaker characteristics) $\bar{\bar{s}}(k)$ from Eq. (1).

transfer function $d^i(k)$ for an exemplary speaker, according to Eq. (3) is presented in the Fig. 2. The final results of the method effectiveness is presented in Fig. 3, according to the accepted indicator of quality $\bar{R}^i$, as the mean, relative reduction of distance between referenced phonemes of different speakers. It is clearly visible that the introduced normalization method shows high efficiency. The phoneme, mean distance reduction reaches values up to 20%. The mean reduction of phonemes distance between each combination of speakers pair is set at around 12%.

CONCLUSIONS AND FUTURE WORK

The conducted experiment proves that the Vocal Tract Normalization should be consider as an universal and comprehensive approach to elimination of variation caused by individual features of speakers. The procedures should be extensive and consider various symptoms and changes in speech signal. Not only vocal tract length is important for modelling speaker specificity but bunch of other features. A proper extraction of this features, combined with proper normalization techniques can cause an essential reduction of distances between phonemes extracted from acoustic speech signal.

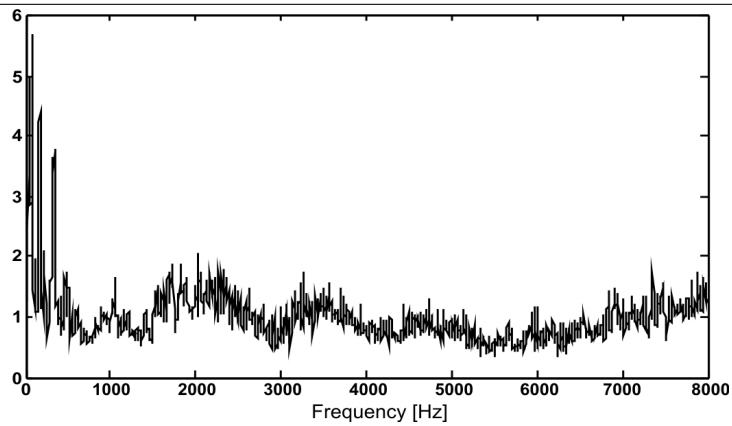

Figure 2. The transfer function $d^i(k)$ for an exemplary speaker, according to Eq. (3).

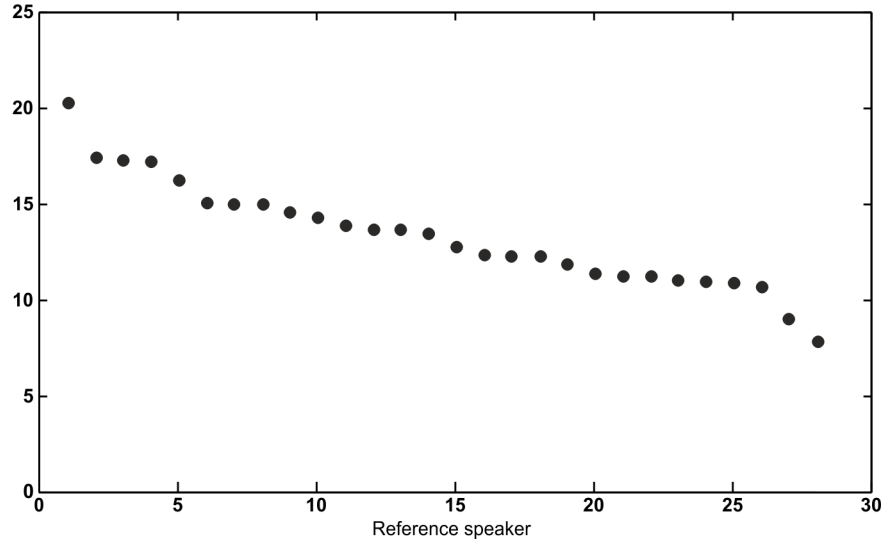

Figure 3. Normalization quality index $\bar{R}^i$, according to Eq. (13).

Our approach can be a useful tool in many speech technology applications. A proper adaptation and implementation of presented procedures can be efficiently included into parameterization process and the performance of speech recognition or classification can be to significantly improve for a relatively low cost of computational time.

The adopted methods of speech normalization verification seem to be a viable alternative for studying the behaviour of the whole speech processing path. However, the further investigation on verification methodology is necessary. All the specificities of speech processing system need to be taken into account. Further considerations are also needed for testing the whole operation on larger recordings corpora, as well as, implementation of normalization, including both, the presented method and algorithm with standard frequency warping.

ACKNOWLEDGMENTS

The project was funded by the National Science Centre granted on the basis of a decision DEC-2011/03/B/ST7/00442.

REFERENCES

[1] P. C. Woodland: *Speaker adaptation for continuous density HMMs: A review* (2001), 11–19, ISCA ITR-Workshop on Adaptation Methods for Speech Recognition.

[2] S. Kacprzak, M. Ziółko, M. Mąsior, M. Igras, and K. Ruszkiewicz: *Statistical analysis of phonemic diversity in languages across the world*, Proceedings of XIX National Conference on Applied Mathematics in Biology and Medicine, 2013.

[3] M. Ziółko, M. Igras, and S. Kacprzak: *Phonemes analysis for genealogical tree of world languages* (2013), Ways to Protolanguage 3 Conference.

[4] H. Boril and J. Hansen: *Unsupervised equalization of Lombard effect for speech recognition in noisy adverse environments*, IEEE Transations on Audio Speech and Language Processing **18** (2010), 1379–1393.

[5] G. Garau, R. Renals, and T. Hain: *Applying vocal tract length normalization to meeting recordings*, Proc. Eurospeech, 2005, pp. 265–268.

[6] J. Lung, M. Salam, M. Rahim, and A. Ahmad: *Implementation of vocal tract length normalization for phoneme recognition on TIMIT speech corpus*, Proc. of the International Conference on Information Communication and Management **16** (2011), 136–140.

[7] P. Moreno, B. Raj, E. Gouvea, and R.M. Stern: *Multivariate Gaussian based cepstral normalization for robust speech recognition* (1995), 137–140, Proc. of the International Conference on Acoustics, Speech and Signal Processing.

[8] F. Mueller and A. Mertins: *Contextual invariant-integration features for improved speaker-independent speech recognition*, Speech Communication **53** (2011), 830–841.

[9] M. Pitz and H. Ney: *Vocal tract normalization equals linear transformation in cepstral space*, IEEE Transactions on Speech and Audio Processing **13** (2005), 930–944.

[10] A. Sarkar, S. Umesh, and S. Rath: *Text-independent speaker identification using vocal tract length normalization for building universal background model* (2009), 2311–2314, Proc. of INTERSPEECH.

[11] D. Sundermann, A. Bonafonte, H. Ney, and H. Hoge: *Time domain vocal tract length normalization* (2004), Proceedings of the ISSPIT.

[12] M. Ziółko, M. Mąsior, B. Ziółko, and M. Igras: *Vocal tract normalisation in computer games* (2013), Proceedings of Signal Processing, Pattern Recognition and Applications.

[13] B. L. McKinley and G. H. Whipple: *Noise model adaptation in model based speech enhancement* (1996), 633–636, Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP-96.

[14] J. O. Smith and P. Gossett: *A flexible sampling-rate conversion method* (1984), 112–115, Proceedings of Int. Conf. Acoustics, Speech, and Signal Processing ICASSP-84.

[15] S. Grocholewski: *CORPORA – Speech Database for Polish Diphones* (1997), 1735–1738, Proc. Eurospeech Conference.

[16] M. Ziółko, J. Gałka, B. Ziółko, and T. Drwięga: *Perceptual wavelet decomposition for speech segmentation* (2010), 2234–2237, Proceedings of the Interspeech Conference.