

Ulrike Glavitsch

Speaker normalization with respect to F_0 : a perceptual approach

*TIK-Report
Nr. 185, December 2003*

Ulrike Glavitsch
Speaker normalization with respect to F_0 : a perceptual approach
December 2003
Version 1
TIK-Report Nr. 185

Computer Engineering and Networks Laboratory,
Swiss Federal Institute of Technology (ETH) Zurich

Institut für Technische Informatik und Kommunikationsnetze,
Eidgenössische Technische Hochschule Zürich

Gloriastrasse 35, ETH Zentrum, CH-8092 Zürich, Switzerland

Speaker normalization with respect to F_0 : a perceptual approach

Ulrike Glavitsch*

December 22, 2003

Abstract

A speaker normalization scheme that uses explicit knowledge of acoustic phonetics is presented. The scheme warps the frequency axis linearly in critical band rate with respect to the fundamental frequency F_0 . It thus allows an immediate adaptation to a new speaker which is an advantage over commonly used schemes. Variants with different values of F_0 and different parameters have been evaluated on several tasks of SpeechDat(II). The results show significant performance improvements on three tasks with monophone models, the most prominent result is a reduction in WER of 44.5 % for an isolated digit task. However, the results achieved with tied triphone models are very modest. It is argued that the normalization scheme may still be correct but that the MFCC feature extraction erases its effect. Evidence for the need of a new feature extraction method that locates spectral peaks and ignores irrelevant portions of the spectrum is given.

Keywords: *speaker normalization, frequency warping, vocal tract length normalization, human speech perception, feature extraction*

1 Introduction

Speaker normalization schemes compensate for variations in speech signals due to different vocal tract sizes of speakers. The most widely used approach performs a linear or piecewise linear warping of the frequency axis. The warping factor is computed per speaker and is estimated by a maximum-likelihood (ML) approach [4]. The ML approach finds the optimum warping factor such that the acoustic models that were trained with features extracted from the warped spectrum show the highest performance. The bilinear transform - an alternate frequency warping scheme - performs a mapping from the unit circle to itself.

*Computer Engineering and Networks Laboratory (TIK), ETH Zürich, CH-8092 Zürich, Switzerland, e-mail: glavitsch@tik.ee.ethz.ch. Partially supported by the Swiss National Science Foundation in the scope of the National Center of Competence in Research (NCCR) on Multimodal Information Management (IM2).

The function parameter is regarded as the speaker’s warping factor and is estimated by an ML approach [7]. Both the linear frequency warping and the bilinear transform have shown significant performance improvements on a number of speech tasks [4], [18], [17], [7]. In contrast to ML-based approaches there exist very few parametric approaches for speaker normalization. An exponential warping function that accomodates for the fact that the high frequencies are more affected by vocal tract changes than low frequencies is presented in [2]. The parameter of the non-linear warping function is computed from measurements of the speaker’s third formant. The average of the third formant is supposed to be directly related to the speaker’s vocal tract length. The effectiveness of the parametric warping function was shown on a single speech database but could not be confirmed on a broader spectrum of speech tasks [2], [18].

The fundamental frequency F_0 plays a central role in human speech perception. The quality of one-formant vowels changes if F_0 is increased or decreased [13]. Listeners hear a different speech sound in these cases. However, information about F_0 is rarely used to improve automatic speech recognition systems [6]. The modulation theory provides a framework that explains how speech is produced and perceived [12]. It suggests a different approach of speaker normalization. According to the theory, the organic information in a speech signal, e.g. the speaker’s age and sex, can be filtered out by a grossly linear warping of the frequency axis in critical band rate (CB). Experimental studies with synthetic vowels have shown that the distance between F_1 (first formant) and F_0 in CB is the most important cue for the perception of vowel openness [10]. This means that the low frequencies, i.e. frequencies around F_1 , are evaluated by humans with respect to F_0 .

This report presents a speaker normalization method that bases on the findings in acoustic phonetics mentioned above. It is a spectrum transformation that performs a linear shift in CB on the full frequency range. The extension to the full frequency range can be made, first, because the modulation theory proposes a grossly linear shift for all frequencies and, second, because the speech data used for evaluation shows little or no vocal effort. In speech with much vocal effort, the low frequencies are raised to a much higher extent than the high frequencies due to the increased F_0 [11]. Several variants of the normalization scheme are presented. They differ in the value of F_0 that is used and in the distance in CB the frequencies are shifted.

The outline of the report is as follows. Section 2 describes the normalization scheme in general with its implemented variants. Section 3 shows the results achieved on several tasks of SpeechDat(II). Section 4 discusses the results. Finally, Section 5 draws conclusions and gives an outlook to future work.

2 Method

2.1 Frequency warping

The normalization scheme performs a frequency warping that shifts frequencies linearly in critical band rate (CB) with respect to F_0 . The spectral shift is performed on the bark

scale that is the most commonly used measure of CB. The normalization transforms the original scale z to the scale z' as follows:

$$z' = z - k(Z_0 - Z_{0,norm}) \quad (1)$$

where

| | | |
|--------------|---|------------------------|
| Z_0 | : | F_0 in CB |
| $Z_{0,norm}$ | : | normalized F_0 in CB |
| k | : | weighting factor |

The value of the normalized F_0 has been set to 120 Hz that corresponds to the pitch of a typical male speaker.

The Hertz-to-bark conversion and inversely is computed by the following expression that has shown to be the most suitable for speech analysis applications [11]:

$$z = \frac{26.81f}{1960 + f} - 0.53 \quad (2)$$

The normalized spectrum is subjected to MFCC feature extraction. 12 cepstral coefficients are computed from the energies of 26 mel filters. Empty or partially filled mel filters that occur at either end of the normalized spectrum are replaced with the closest mel filter that is at least 50 % full. This measure fills up empty and partially filled mel filters in a probable way. The zeroth cepstral coefficient is added to the feature vector as the 13th component. Finally, delta and acceleration coefficients are generated to build the final feature vector of length 39.

2.2 Estimation of F_0

Equation 1 uses F_0 in critical band rate (CB). The most commonly known value of F_0 is the instantaneous F_0 . The instantaneous F_0 exists in voiced sections of the speech signal whereas it is undefined or set to a default value in the voiceless segments. Pitch tracking algorithm decide on the presence of voice at any position of the speech sequence and compute the instantaneous F_0 within the voiced sections. However, humans seem to evaluate the spectrum, more precisely, the spectral peaks shaped by the formants with respect to the base value of F_0 [14]. The base value of F_0 (prosodic baseline) is significantly lower than the instantaneous F_0 and roughly follows the local minima of the instantaneous F_0 curve. The normalization scheme is evaluated with both types of F_0 values, thus, algorithms for the computation of either one are required.

The instantaneous F_0 is computed with the standard software *ESPS-get_f0* that implements a standard pitch tracking algorithm [9]. The instantaneous F_0 is gained in two passes in order to get a smooth curve with as few outliers as possible. The first pass computes an estimate of the instantaneous F_0 using the default parameters of *ESPS-get_f0*. An average F_0 is calculated from this first curve denoted as $\overline{F_0}$. In the second pass, the parameters of *ESPS-get_f0* are set such that the maximum instantaneous F_0 must not exceed $1.5\overline{F_0}$

and the minimum allowable instantaneous F_0 must not fall below $0.5\overline{F_0}$ assuming that a speaker’s pitch varies within an octave in neutral speech.

There are no algorithms available for the computation of the base value of F_0 , but an estimation can be derived from findings in human speech perception. Decisions about the F_0 value do not seem to be made instantaneously. Humans rather let a certain period of time pass before they decide on the F_0 value to be used. This period of time has been observed to be the length of a syllable or roughly 400 ms [12]. Thus, the following estimation is proposed. The base value of F_0 is computed as the lowest value of the instantaneous F_0 curve within the last 400 ms. The base value of F_0 is set to a default value, i.e. 0, if the interval of 400 ms is unvoiced indicating that the base value of F_0 does not exist at this position. In all other cases, the base value of F_0 is set to the minimum F_0 value of the voiced portions within this interval. In case the speech sequence is less than 400 ms long the computation is performed on the given length of the sequence.

This way of F_0 base value estimation has the following implications:

- The base value of F_0 is defined for unvoiced speech segments with a left voiced context as long as the unvoiced section is less than 400 ms long.
- The base value of F_0 is not defined for unvoiced phonemes at the beginning of an utterance as there exists no previous instantaneous F_0 that could be used at this position.

A study in acoustic phonetics shows that fricatives in voiced context are influenced by their voiced context [3]. The phoneme boundaries of sibilants change if they occur in voiced context. It is suggested that sibilants in context need to be normalized but not to such a large extent as vowels. The normalization scheme uses equally weighted shifts at all positions where the base value of F_0 is defined.

A number of speech segments, their corresponding instantaneous F_0 curve and the estimated base values of F_0 are shown in Fig. 1 and Fig. 2. The example in Fig. 1 shows that the base value of F_0 is less excursive than the instantaneous F_0 . The minimum and maximum value of the instantaneous F_0 are 114 and 297 Hz whereas they are 114 and 240 Hz for the base value of F_0 . It can also be observed that the base value is well defined for the unvoiced region in the middle of the utterance. The base value of F_0 exists also for the portion of the signal after the last voiced section. Fig. 2 shows an example that both starts and ends with a fricative. The base value of F_0 at the beginning of the fricative is not defined whereas it exists for the fricative at the end of the utterance. This can be seen as follows. The curve of both the instantaneous F_0 and the base value of F_0 start at the same position, i.e. at the beginning of the voiced section that is the beginning of the vowel /ue/. Thus, the base value of F_0 for the fricative at the beginning of the utterance is not defined. However, the base value of F_0 is defined for a much longer period than the instantaneous F_0 that ends after the voiced /n/ of the utterance. The fricative /f/ falls exactly in this period as it follows the voiced portion of the utterance directly.

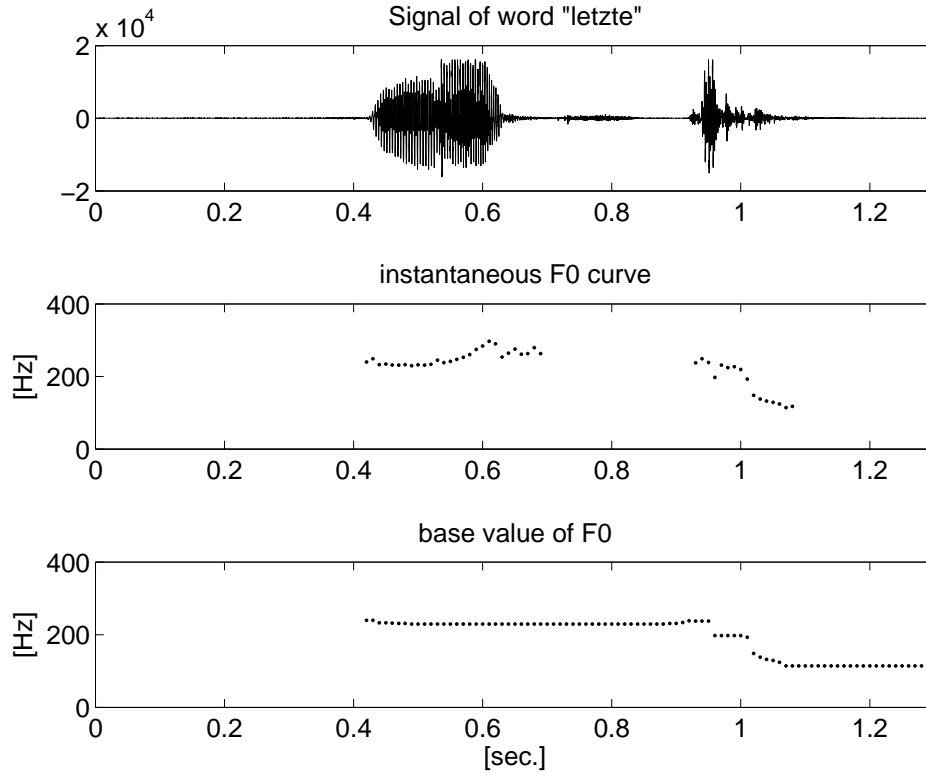


Figure 1: Speech signal, instantaneous F_0 and base value of F_0 of the German utterance 'letzte' (English translation 'last')

2.3 Variants of the normalization

The normalization scheme was implemented with two different F_0 values, the instantaneous F_0 and the base value of F_0 . As a result, variant (I) that uses the instantaneous F_0 and variant (II) with the base value of F_0 existed. The subvariants (IIa) - (IId) were created to explore different weighting factors k of the normalization shift. Given the importance of the base value of F_0 in human speech perception no subvariants of variant (I) were implemented. The 5 variants of the normalization scheme are listed in the following:

- (I) normalization with respect to the instantaneous F_0 , $k = 1.0$.
- (IIa) normalization with respect to the base value of F_0 , $k = 1.0$.
- (IIb) do., $k = 0.7$.
- (IIc) do., $k = 0.6$.
- (IIId) do., $k = 0.5$.

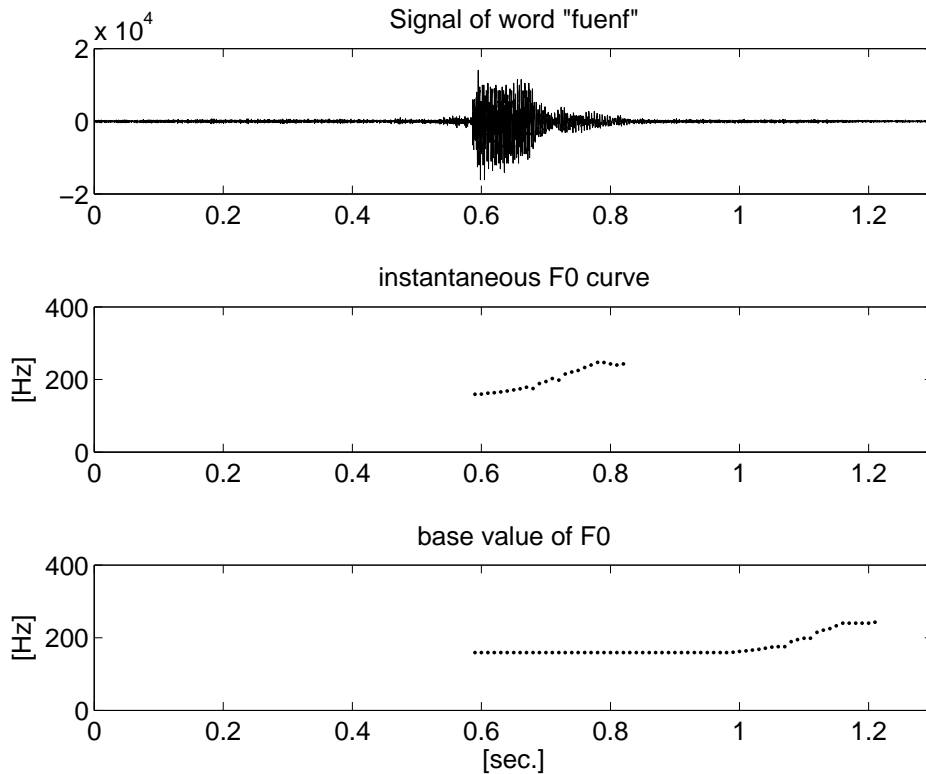


Figure 2: Speech signal, instantaneous F_0 and base value of F_0 of the German utterance 'fünf' (English translation 'five')

A weighting factor of 0.63 can be derived from the regression equation for vowel openness in [16]. Thus, it is natural to explore the weighting factors 0.7, 0.6, 0.5 in variants (IIb) - (IIId).

3 Experiments

Variants (I) - (IIId) of the normalization scheme were evaluated on SpeechDat(II) - a standardized speech corpus described in [5]. The Swiss German database of SpeechDat(II) used here contains speech sessions of 2000 different speakers. Sessions of 1500 speakers are used for training and sessions of 500 speakers for testing. Two different classes of HMMs were created during training: monophones and tied triphone models. Model sets with 1, 2, 4, 8, 16, and 32 mixtures were trained for both classes.

The variants of the normalization scheme were evaluated on 5 speech tasks of SpeechDat(II). The selected tasks are shown in Tab. 1. The variants (I) - (IIId) were evaluated against the reference recognition system that uses no normalization. It is referred to as the baseline system.

The recognition results of the baseline system and variants (I) - (IIId) achieved on each of

| test | recognition task |
|------|-------------------------------|
| Q | ja/nein (yes/no) answer words |
| I | isolated digits |
| A | 30 isolated application words |
| O | city names |
| W | phonetically rich words |

Table 1: Selected test tasks for the evaluation of the normalization scheme.

| | Q | I | A | O | W |
|----------|------------|-------------|-------------|-------|-------|
| baseline | 0.22 | 3.71 | 5.25 | 16.81 | 42.93 |
| (I) | 0.54 | 3.51 | 5.34 | 19.15 | 43.96 |
| (IIa) | 0.33 | 4.74 | 6.22 | 20.64 | 45.44 |
| (IIb) | 0.0 | 2.47 | 5.15 | 19.15 | 45.60 |
| (IIc) | 0.0 | 2.68 | 5.15 | 19.15 | 45.50 |
| (IId) | 0.0 | 2.06 | 4.96 | 17.87 | 44.39 |

Table 2: Performance in % WER achieved with monophone models.

the selected tasks are shown in Tab. 2 and Tab. 3. Tab. 2 shows the lowest word error rate (WER) in % achieved on the set of monophone models for each variant and the baseline system for each of the selected tasks. The figures in Tab. 3 represent the lowest WER in % achieved on the set of tied triphone models. Some of the figures in Tab. 2 and Tab. 3 are typed in boldface letters. They represent lower WERs than the corresponding WER of the baseline system.

4 Discussion

The new normalization scheme shows lower word error rates than the baseline system on three of the selected tasks with monophones but only on a single task with tied triphone models.

Variants (IIb), (IIc), and (IId), i.e. the variants that use the base value of F_0 and weighting factors 0.7, 0.6, and 0.5, outperform the baseline system on tasks Q, I, and A with monophones. On task I, the reductions in WER are 33.4 % for variant (IIb), 22.8 % for variant (IIc), and 44.5 % for variant (IId). This performance improvement is comparable or even higher than that of a standard speaker normalization scheme [4]. On task A, variants (IIb) and (IIc) achieve a reduction in WER of 1.9 % over the baseline system, the reduction in WER for variant (IId) is 5.5 %. The WERs for task Q are very small and the differences among them are statistically not relevant. Thus, the reductions in WER are not given.

The WERs on tasks O and W are higher than those of the baseline system for all 5

| | Q | I | A | O | W |
|----------|------|------|-------------|-------|-------|
| baseline | 0.22 | 1.03 | 0.58 | 9.17 | 25.05 |
| (I) | 0.43 | 1.65 | 0.68 | 10.43 | 27.60 |
| (IIa) | 0.33 | 2.27 | 1.07 | 12.34 | 28.97 |
| (IIb) | 0.43 | 1.65 | 0.58 | 9.57 | 26.64 |
| (IIc) | 0.22 | 1.44 | 0.39 | 10.00 | 26.06 |
| (IId) | 0.22 | 1.44 | 0.68 | 9.79 | 25.37 |

Table 3: Performance in % WER achieved with tied triphone models.

variants of the normalization scheme with monophones.

On the tied models, variant (IIc) only, i.e. the variant using the base value of F_0 and a weighting factor of 0.6, achieves a lower WER than the baseline system. The performance improvement is gained on a single task, namely task A. All other normalization variants perform equally or worse than the baseline system.

It has to be noted that variant (I) that uses the instantaneous F_0 shows lower WERs on all tasks except for task Q both with monophones and tied triphone models than variant (II) that uses the base value of F_0 . On task I, variant (I) even outperforms the baseline system with monophones.

5 Conclusions and outlook

Several variants of a normalization scheme that shift the spectrum linearly in critical band rate with respect to the fundamental frequency F_0 have been evaluated on several tasks of SpeechDat(II). The variants that use a weighting factor of 0.5 - 0.7 outperformed the baseline system on three of the five test tasks if monophone models were used. However, the results achieved with tied triphone models are very modest for almost all variants of the scheme. The author refrains from interpreting the results in more detail here but rather tries to put them in a more global picture.

The following conclusions can be drawn from the results achieved with the presented normalization scheme:

1. The normalization is not correct yet.
2. The normalization is correct in principle but its effects are undone by the MFCC feature extraction.

Besides the fact that the presented normalization scheme can still be improved the author is inclined to adhere to interpretation 2. Why?

In acoustic phonetics it is commonly accepted that stationary phonemes are characterized by the position of the first few formants and F_0 . Spectral peaks constitute the information-conveying parts of the spectrum [15], [8]. This means that only small portions

of the spectrum contain relevant information whereas the remaining parts are of less or no importance.

Standard MFCC features represent the gross shape of the short-time spectrum. The energies in critical bands - so-called mel-filters - are computed and a cepstral analysis on the mel-filtered energies is performed from which only the n lowest cepstral coefficients are used as the final features. This way, the full spectrum is used by the feature extraction and both relevant and irrelevant information is included in the mel-filtered energies. In addition, the computation of mel filters often leads to a loss of information on where spectral peaks are located, particularly, in the case of neighboring spectral peaks. In this case, the mel filters between the spectral peaks may have a higher energy than those at the peak positions. Both the fact that MFCCs contain a lot of insignificant information and that they do not locate spectral peaks precisely may be the reason why current parametric speaker normalization approaches are not successful (the presented normalization scheme is also a parametric approach). The only successful speaker normalization schemes are those that compute the warping function based on a maximum-likelihood (ML) approach as mentioned in Section 1. The ML approach, however, computes the parameter of the warping function in such a way that the effect of the insufficiencies of MFCC feature extraction is minimized whereas the effect of its strength is maximized.

The author claims that the validity of the presented normalization scheme cannot be shown with standard MFCC features. In fact, a new feature extraction method needs to be developed that locates peaks in the spectrum by circumventing the formant estimation problem. Formants are almost impossible to estimate in high-pitched speech because of the highly undersampled spectrum. The problem of locating spectral peaks can be reduced to a pattern matching problem using the approach suggested by de Cheveigné [1].

Future work will focus on the following: (1) the development of a new feature extraction method that detects peaks in the spectrum as suggested above and (2) the optimization of some parts of the normalization scheme. What the optimization of the normalization scheme is concerned the estimation of the base value of F_0 needs to be refined. For instance, the asymmetry that the base value of F_0 of an unvoiced speech segment at the beginning of an utterance is undefined whereas it is defined for a voiceless section at the end of an utterance is not satisfying yet. An idea might be to not only consider a time frame of the past instantaneous F_0 values but also look a time frame in the future to compute the base value of F_0 .

References

- [1] A. de Cheveigné and H. Kawahara. Missing-data model of vowel identification. *Journal of the Acoustical Society of America*, 105(6):3497–3508, 1999.
- [2] E. Eide and H. Gish. A parametric approach to vocal tract length normalization. In IEEE, editor, *Proceedings of the ICASSP*, pages 346–348, 1996.
- [3] T. Krull, D. An experiment on the cues to the identification of fricatives. In *Proceedings of the 11th ICPHS*, volume 5, pages 205–208, 1987.

- [4] L. Lee and R. C. Rose. Speaker normalization using efficient frequency warping procedures. In *Proceedings of the ICASSP*, pages 353–356. IEEE, 1996.
- [5] B. Lindberg, F. T. Johansen, N. Warakagoda, G. Lehtinen, et al. A Noise Robust Multilingual Reference Recogniser Based on SpeechDat(II). In *Proceedings of the ICSLP'2000*, Beijing (China), October 2000.
- [6] M. Magimai-Doss, T. A. Stephenson, H. Bourlard. Using pitch frequency information in speech recognition. In *Proceedings of the Eurospeech*, 2003.
- [7] J. McDonough, W. Byrne, X. Luo. Speaker normalization with all-pass transforms. In *Proceedings of the ICASSP*. IEEE, 1999.
- [8] M. R. Schröder. Linear prediction, extremal entropy and prior information in speech signal analysis and synthesis. *Speech Communication*, 1:9–20, 1982.
- [9] D. Talkin. A Robust Algorithm for Pitch Tracking (RAPT). In Kleijn and Paliwal, editors, *Speech Coding and Synthesis*, chapter 20, pages 495–518. Elsevier, 1995.
- [10] H. Traunmüller. Perceptual dimension of openness in vowels. *Journal of the Acoustical Society of America*, 69(5):1465–1475, May 1981.
- [11] H. Traunmüller. Paralinguistic Variation and Invariance in the Characteristic Frequencies of Vowels. *Phonetica*, 45:1–29, 1988.
- [12] H. Traunmüller. Conventional, biological and environmental factors in speech communication: A Modulation Theory. *Phonetica*, 51:170–183, 1994.
- [13] H. Traunmüller. The role of F_0 in vowel perception. <http://www.ling.su.se/staff/hartmut/i.htm>, 1998.
- [14] H. Traunmüller. Evidence for demodulation in speech perception. In *Proceedings of the 6th ICSLP*, volume III, pages 790–793, 2000.
- [15] H. Traunmüller and F. Lacerda. Perceptual relativity in identification of two-formant vowels. *Speech Communication*, 6:143–157, 1987.
- [16] H. Traunmüller, A. Eriksson, L. Ménard. Perception of speaker age, sex, vocal effort and vowel quality investigated using stimuli produced with an articulatory model. In *Proceedings of the XVth ICPHS*, pages 833–836, 2003.
- [17] P. Zhan and A. Waibel. Vocal Tract Length Normalization for Large Vocabulary Continuous Speech Recognition. Technical Report CMU-CS-97-148, School of Computer Science, Carnegie Mellon University, 1997.
- [18] P. Zhan and M. Westphal. Speaker normalization based on frequency warping. In *Proceedings of the ICASSP*. IEEE, 1997.