VTLN through Frequency Warping Based on Pitch

Carla Lopes[1,2], Fernando Perdigão[1,3]

[1]Institute of Telecommunications
Pólo II, FCTUC, Pinhal de Marrocos, 3030-290 Coimbra, Portugal

[2]Polytechnique Institute of Leiria-ESTG, [3]University of Coimbra-DEEUC
{calopes, fp}@co.it.pt

Abstract - **This article describes a vocal tract length normalization (VTLN) procedure through pitch based frequency warping. This procedure aim to reduce de inter-speaker variability, present in speech signals. It is also described a method for coarticulation phenomena compensation, that reduce speech signal variability due to phonetic context. This procedure operates at the phonetic level since makes a modeling of coarticulation events, and at the linguistic level since these units lead to alternative pronunciation rules.**

Inter speaker variability removal is performed by a traditional speaker normalization method, which consists in expanding or compressing the *Mel* bank filter bandwidth in order to normalize the vocal tract length (VTL) of each speaker to a standard one. The estimation of VTL is, in previous works, based on formant information, but authors pointed out as an obstacle for better results, the difficulty of formant frequency estimation. The method presented in this paper overcomes such problem since we estimate the warping factor (WF) through pitch. The recognition results presented on this paper, for a telephone digit recognition task prove that this procedure leads to similar improvements to those obtained with traditional methods based on formant information.

I. Introduction

In speaker independent automatic speech recognition (ASR) systems, speech models are trained making use of a great amount of speech, pronounced by a great variety of speakers as well. Each speaker has specific features, which are not only related with physiological features, as length and shape of vocal tract, but also with linguistic aspects such accent, dialect, stress and environment. Due to these speaker specific features, and due to the differences between all speakers, speech signals arise to the system with different acoustic features, resulting on models in which spectral distributions usually have great variances and therefore great overlapping through distinct phonemes. This situation is an obstacle for ASR system performance.

The effects caused by the variability of speech, in addition to the adverse conditions of the environment, make the greatest challenge for actual speaker independent ASR systems. Generally, the sources of variability related with the speakers can not be totally eliminated. It is therefore necessary that ASR technology model efficiently this kind of obstacle.

The present work is based on an efficient recognition of connected digit strings. The process of speaker independent recognition of connected digits through the telephone, is a special and an interesting case for ASR. On one hand this is a relative simple task, since the vocabulary involved is small, on the other hand the system must be extremely accurate since one wrong digit on a string result on an invalid string.

To improve the distinction between recognition models we implemented a method for coarticulation phenomena compensation. This compensation is described on section II. To reduce the variances of the spectral distributions of models, produced by speakers with different features, it was implemented a normalization method, described on section III and IV. On section V it is reported the results obtained with the implemented method.

II. Coarticulation Modeling

The speech production mechanism is a temporal set of articulatory actions to produce the sequence of phonemes. For consecutive phonemes these actions overlap, leading to coarticulation. In this situation, the vocal tract is articulating a phoneme at the same time that is preparing the articulation of the next one. So, the degree of coarticulation is highly dependent of the pronunciation rate. The acoustic realization of a set of sounds is greatly related with the fact that our articulators can not move instantaneously from one position to another, leading to that some phonemes are only partially articulated.

The coarticulation is defined by Kirchhoff and Bilmes, [8] as "a change in the acoustic-phonetic content of a speech segment due to the anticipation or preservation of adjacent segments". According to these authors the degree of coarticulation varies with several factors: with the speaking rate; with the degree of syllabic stress and with the quality of the vowel (central/peripheral and strong/weak). The authors concluded that a highly speaking rate associated with a low degree of stress leads to a strong coarticulation.

In our work, we performed a study on sentence sonograms to define the most frequent coarticulation units. We considered the existence of coarticulation when there were a clear continuity of the formants from one digit to the neighbour. For example when the sequence *dojS dojS* ("two two" in Portuguese) is pronounced with no coarticulation, we have a sequence of four models (/doj/, /S/, /doj/, /S/). In the production of the diphthong /oj/ there is a lowering of the second formant followed for a growth. The phoneme */S/* is characterized by an energy cloud. Since */S/* is unvoiced there is no trajectory for F2 in this period. After that appears again a lowering followed for a growth of F2 (due to /oj/) and a new /S/. When *dojS dojS* is pronounced coarticulating the first */S/* coarticulates with the beginning of the adjacent *dojS.* The trajectory of F2 for the first /oj/ remains but there is a perturbation of the */S/* of the first *dojS*. In this period it is clear a continuity of F2 from de first /oj/ to the second one, what make us conclude that there was no pronunciation of /S/. Indeed, the fact that following an */S/* exists a voiced consonant conduct to */S/* appears like a */Z/*.

On our sentence sonograms study we found many coarticulation phenomena. Other example appears on the presence of two adjacent vowels, which results on the decay of one of them. This is called elision.

When the speech production has an absence of a clear separation between the acoustic specific features of each phoneme or sub word, probably coarticulation is present. This situation makes more difficult the estimation of these units because the *frames* correspondent to each model won't be rightly attributed. This contributes to the construction of models with a high variance. To avoid this problem we decided to add to our set of models four units of coarticulation. These units are not the only existent, but there are some that occur in a number that allows us an accurate estimation. The most common coarticulation units, and for that the selected ones are presented in table I.

The phoneme /z/ appears when /S/ is followed by a vowel, while /Z/ is when /S/ is followed by a voiced consonant. /u-u~/ or /u-ojt/ are other cases of common coarticulations.

Table I - Coarticulation units selected.

Coarticulated phones	Resulting Phones
S……u~	*z*
S……ojt	*z*
S……zEr	*Z*
S……doj	*Z*
u……u~	u-u~
u……ojt	u-ojt

Since the pronunciation varies, if not considered, the performance of systems degrades. Explicitly modeling these pronunciation variations we could correct some errors induced by speakers production variability.

The system acts at two levels: on the acoustic level since coarticulation models were introduced and on the phonological level since these new models originate alternative rules of pronunciation.

III. SPEAKER NORMALIZATION PROCEDURE

Attempting to reduce the speech signal variability (due to inter speaker differences) and produce significant improvements on ASR performance, different techniques have been investigated to normalize the parametric representation of speech signals through the manipulation of its acoustic parameters. One of the techniques, widely used in speaker normalization, is the frequency warping axis technique. This technique appears as an attempt to normalize the vocal tract length (VTL) of different speakers, reducing their influence on the spectral parameters. Using this method, the acoustic parameters are transformed by warping the speech signal in the frequency domain. This warping can be performed in two distinct ways. The first, which is proposed by [1],[11],[12] is performed by compressing or expanding the speech signal, in the spectral Fourier domain. The second one, proposed by [2],[13] is performed by the compression or expansion of the filter bank responses, used in MFCCs (*Mel* Frequency *Cepstral* Coefficients) estimation, in the *Mel* scale. Whether the warping is applied on the spectral signal or directly on the filter bank the goal is similar: both attempt to map the spectrum of a phoneme pronounced by distinct speakers to a standard one. This mapping is performed by a warping function that depends on a single warping factor. The selection of this parameter and the shape of the function is vital for the application success. With regard to the shape a wide variety of functions were proposed: linear like the work of Lee and Rose, [7]; piecewise linear as in Wegmann et all, [11]; non linear like Eide and Gish, [1] or bilinear as in the case of Zhan and Waibel, [13] and Fukada and Sagisaka, [2].

With regard to the selection of the warping factor there are two main procedures: the selection based on maximization of likelihood (ML), [7],[11],[13] and the selection based on speaker specific acoustic parameters, [1],[4]. The first one uses a predefined set of warping factors and, following an iterative procedure based on ML, selects the best WF for a specific speaker. The WF is selected so that the probability of a set of acoustic features (of a given speaker) is maximized in regard to an acoustic model taken as reference. The second one selects the WF using an approach based on the measurement of the frequencies of the formants of the speaker, since according to the authors the position of these reflects the VTL.

Several authors obtained better results using ML criterion, however, the method based on speaker specific parameters has the important advantage of being computationally less expensive.

However the estimation of formants is liable to errors, especially when the system works in adverse conditions.

In our work, and to overcome such problems we selected the warping factor from pitch (F0).

IV. Pitch Based Frequency Warping

It is somewhat intuitive that the estimation of the VTL is supported by acoustic studies. However, as already reported this directly estimation from the speech signal is difficult since there isn't a simple relationship between the formants and VTL.

To deal with this situation Eide and Gish, [1] proposed a method, which score conduct to significant improvements in performance. Their proposal is based on the warping function given by the equation (1) and sketched in figure 1, where k_s is the ratio between the median third formant (F3) of a given speaker and the median of F3 of all speakers of the train set.

$$f' = k_s^{\frac{3f}{8000}} f \qquad (1)$$

The preference on F3 is due to the fact that this formant is the more stable, i.e., is less dependent of linguistic information and therefore from the statistic point of view is the more robust. Zhan and Westphal [12] also defended this point. They expand the work of Eide and Gish, [1] and make the normalization using the same warping function but estimating the warping factor also from the 1st and 2nd formants, not achieving better results with these formants.

Our method makes use each speaker's pitch to estimate his vocal tract length, and perform normalization. This procedure seems profitable to us since pitch is more stable than F3 and it estimation is more reliable. Pitch determination obliges to a voiced/unvoiced separation, but this is also necessary on formant determination since it does not make sense to estimate formants in unvoiced *frames*.

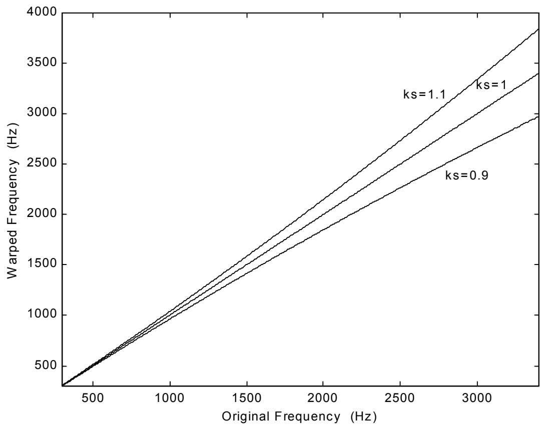

Fig. 1. - WF shape used for the determination of the low and high frequencies of the *Mel* scale filter bank.

Since F3, when used as an indicator of normalization leads to good results and since in our method we intend to perform normalisation from F0 it necessary to analyse if there is any relationship between these two features.

To do so we calculated F0 and F3 for each sentence. The pitch was estimated through the algorithm AMPEX [5], and formants through SFS, [10].

We made a scatter plot between the means and medians of F0 and F3 and found a correlation factor of 0.45 and 0.3 respectively. Due to the fact this last value be significantly below the first one we decided to use means (instead of medians) as a determinative factor for normalization.

In figure 2 it is presented the relation between F0 and F3 means of 2454 digit strings (train and test set).

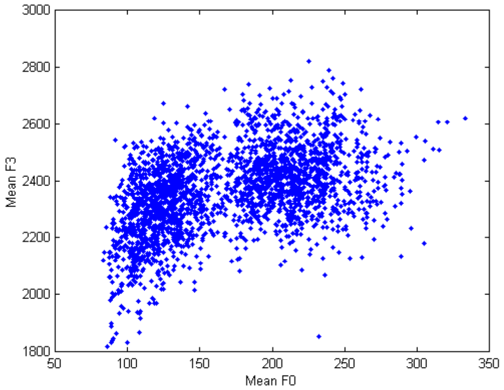

Fig. 2: Scatter plot between means of F0 and means of F3.

We observe that figure 2 seems to be composed by to clouds, each one representing a gender. The left cloud corresponds to sentences of male speakers, since F0 values are under the *standard* speaker's (160Hz), and the right one of female speakers for the opposite reason. It is therefore expected than F0 and F3 distributions follow a bimodal distribution with each mode corresponding to each gender.

On our work we use Eide and Gish's function but defined k_s with F0 instead F3. Since the relation between $\frac{F3}{\overline{F3}}$ and $\frac{F0}{\overline{F0}}$ are different, k_s will be affected by a value that will be given by the expression (2), where $\overline{F0}$ is the mean $F0$ among all speakers.

$$k_S = \alpha \frac{F0}{\overline{F0}} \qquad (2)$$

The question is how to find α. Independently of the speaker gender, each speaker has its own warping factor. But, since one of the major factors of variability is due to gender differences, we decided analyse two sets of speakers, one of each gender, separately.

We computed warping factors for both sets and on two distinct ways. On the first one, WF was obtained as the ratio between the mean of the third formant (F3) of a speaker and the mean of F3 of all speakers in the train set (2300Hz). On the second one k_s is the ratio between the mean of pitch of the voiced set of a given speaker and the mean of the pitch of a standard speaker.

The means were estimated using only voiced frames, detected by a pitch detector.

For both cases the distribution of the warping factors of the female set are above the male set one. This allows us to conclude that F3 frequencies of male speakers are under the female frequencies and obviously the same happens with pitch. However, this last distance is greater. We verified that the mean of pitch of the female set is about 80Hz above the mean pitch of the male set.

Comparing both distributions we verified that they have similar shapes. This led us to implement a mapping between the two distributions.

If the mapping were made according to the speaker gender, we can easily find a set of points that make a correspondence between F0 and F3, and find a function that maps the pitch of a speaker on his F3. However our method intend to normalize speakers besides its gender. So we analyze separately the gender's distributions, found a set of points (one to each gender) that maps the distribution of F0 on F3.

To obtain a mapping function gender independent, we considered that under the *standard* speaker's pitch (160Hz) the mapping refers to male speakers and the above from female speakers and made a 3rd order fitting on these points. This mapping function is drawn in figure 3. The ordinate $g(F0)$ establishes the value that F0 should be warped to agree with the F3 value. This function establishes α of the equation (2), that is $\alpha = g(F0)$.

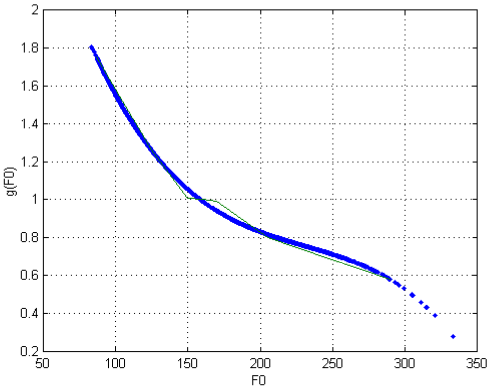

Fig. 3. *g* function that establishes α of equation (2).

The normalization is performed in three distinct steps. The first one consists on the determination of the mean of pitch in voiced frames. Then we calculate the warping factor through the equation (2), with α given by the function $g(F0)$ of figure 3. Finally, we change the upper and lower frequencies of the *Mel* scale filter bank through the function given by equation (1).

The warping function proposed is then:

$$f' = \left(g\left(F0\right) \frac{F0}{\overline{F0}} \right)^{\frac{3f}{8000}} \times f \qquad (3)$$

To not exceed the *Nyquist* frequency the warping factor were restrict to an interval from 0.9 to 1.1.

V. Experiences

The tests of this work were performed using strings of nine connected digits presents on a Portuguese speech database, collected through the telephone network called TELEFALA, [9]. The speech signals were recorded at a sampling frequency of 8kHz and formatted with PCM-μ law. The training set has 1012 digits strings and the test set has 847. The number of female and male speakers is about 50% in both sets. Since each speaker utters a reduced number of utterances, there is a high variability inter speaker, not only related with physiological differences but also differences related with geographic origins, rhythm and style of production. The both sets were labeled manually, based on phones and sub word units. The choice of these units was based on an acoustic-phonetic study of Portuguese digits. We found units acoustically well characterized, which correspond mostly to syllables or phonemes of the digits. Since there are much coarticulation between adjacent digits there were considered four coarticulation units, described in section II. Analyzing the utterances of the database we found that exists, beyond digits, other occurrences of non linguistic events that was labeled as well. 18 phonemes and 4 coarticulation units were defined to describe the 10 digits and models for these units were trained. Additionally, we considered 10 other models for capture the statistical properties of the silence, noise and out of vocabulary words.

To model the phonemes we used continuous Hidden Markov Models (HMMs) with left to right topology. The train was performed using HTK 2.1,[14] software.

The acoustic parameters used were MFCCs, with energy and the corresponding delta coefficients.

The recognition results are sex independent and are presented in table II and correspond to models with 2 and 8 components of the Gaussian mixture for each HMM state.

We named *Coart Models* to the experiences that contain coarticulation models. We evaluate the different procedures by comparing WER (Word Error Rate) and SER (Sentence Error Rate) values. The improvements presented are related to each test when comparing with the baseline.

Table II - Recognition Rates.

System	**Mixture number**	**WER**	**SER**	**WER Improvements (comparing with baseline)**	**SER Improvements (comparing with baseline)**
Baseline	2	7,1%	39,0%		
Baseline	8	4,5%	27,9%		
Coart. Models	2	6,7%	37,7%	6,9%	3,5%
Coart. Models	8	3,9%	24,1%	**13,8%**	**15,7%**
Normalization based on F3	2	6,5%	36,7%	9,9%	6,1%
Normalization based on F3	8	3,8%	23,1%	**18,7%**	**20,4%**
Normalization based on F0	2	5,6%	31,6%	28,1%	23,1%
Normalization based on F0	8	3,7%	22,1%	**20,0%**	**26,2%**

Modeling coarticulation events we achieved improvements, with 2 Gaussian mixtures of 6.9% in WER and 3.5% in SER and with 8 mixtures the results were 13.8% in WER and 15.7% in SER.

In what concerns to the normalization procedure, our method not only reaches the results of the function proposed by Eide e Gish [1] (normalization based on F3) as outperformed them. The results of complete utterance, with 2 mixtures show an improvement of 23.1% for our method and 6.1% to Eide´s method faced to the baseline. With 8 mixtures the results are even better, 26.2% for normalization based on F0 e 20.4% for based on F3. However the results with increased number of mixtures did not accomplish the previous, leading us to conclude that it will be necessary a fewer number of Gaussian mixtures to model each sub word. Since the results of recognition were superior it is expected that the sub words models became more compact.

Although it is not present in table 2, the best results were obtained with 17 Gaussian mixtures, with normalization based on pitch. The digit recognition result was 96.9% and the sentence recognition result was 81.6%.

Additionally we tested the method considering models of entire word digits. In this case the results did not evidence improvements over 0.8% WER and 6.6% SER obtained with 20 mixtures.

VI. Conclusions

The proposed normalization method, which is based on pitch, proved to be of great utility in the improvement of performance of a 9 connected digit string task. The method overcomes the dependency of the system performance face to the reliability of formant estimation.

Gouvêa, [3], in his work, uses median of de tree formants. He pointed out that the system performance only stabilizes when each speaker data reaches 12s. With our system we get a reasonable estimation of pitch after a small set of voiced *frames*.

We also believe that formant estimation degradation is higher in noisy conditions than pitch estimation.

The normalization based on pitch reached the results obtained with formants. An improvement of about 26% was achieved over the baseline performance with 8 mixtures. These results point that must exist a good correlation between pitch and vocal tract length. This was the idea that motivated this work. Female speakers have shorter vocal tracts, therefore higher formants, and also higher pitch frequencies.

VII. References

[1] Eide, E., Gish, H., *A Parametric Approach to Vocal Tract Normalization*, Proc. ICASSP' 96, v.1, pp. 346-348, May 1996.

[2] Fukada, T., Sagisaka, Y., *Speaker Normalized Acoustic Modeling Based on 3-D Viterbi Decoding*, Proc. ICASSP'98, v.l. 1, pp. 437-440, Seattle, WA, May 1998.

[3] Gouvêa, E., *Acoustic-Feature-Based Frequency Warping for Speaker Normalization,* Ph.D. Thesis, Carnegie Mellon University, Pittsburgh, December 1998.

[4] Gouvêa, E., Stern, R., *Speaker Normalization Through Formant-Based Warping of the Frequency Scale*, Proc. Eurospeech, Rhodes, 1997.

[5] Immerseel, L., Martens, J., *Pitch and Voiced/Unvoiced Determination with an Auditory Model*, JASA 91(6), pp 3511-3526, June 1992.

[6] Junqua, J, Haton, J., *Robustness in Automatic Speech Recognition*, Kluwer Academic Publishers, 1996.

[7] Lee, L. and Rose, R. C., *Speaker Normalization using Efficient Frequency Warping Procedures*, Proc. ICASSP'96, May 1996.

[8] Kirchhoff, K., Bilmes, J., *Statistical Acoustic Indications of Coarticulation*, Proceedings of ICPh99, pg. 1729-1732, San Francisco, 1999.

[9] Neves, F., Amaral, R., Plácido, P., Marta, E., Perdigão, F., Sá, L., *A Portuguese Telephone Speech Database Collected Using an Automated System*, 7ª Conf. da Associação Portuguesa de Reconhecimento de Padrões, Aveiro, 1995.

[10] Speech Filing System, SFS Release 4.25, Version 1.25, Mark Huckvale, University College of London.
http://www.phon.ucl.ac.uk/resource/sfs/help/index.htm

[11] Wegmann, S., McAllaster, D., Orloff, J., Peskins, B, *Speaker Normalization on Conversational Telephone Speech*, Proc. ICASSP '96, v.1, pp. 339-341, May 1996.

[12] Zhan P., Westphal M., *Speaker Normalization Based on Frequency Warping*, Proc. ICASSP '97, pp. 1039-1042, Munich, Germany, April 1997.

[13] Zhan, P., Waibel, A., *Vocal Tract Length Normalization for Large Vocabulary Continuous Speech Recognition*, CMU-LTI-97-150, May 1997.

[14] Young, S., Jansen, J., Odell, J.,Ollasson D. and Woodland, P., *The HTK Book (for HTK Version 2.1)*, Cambridge University, Cambridge, UK, 1995.