Applying Vocal Tract Length Normalization to Meeting Recordings

Giulia Garau, Steve Renals

Centre for Speech Technology Research
University of Edinburgh
Edinburgh EH8 9LW, UK
{g.garau,s.renals}@ed.ac.uk

Thomas Hain

Department of Computer Science
University of Sheffield
Sheffield S1 4DP, UK
t.hain@dcs.shef.ac.uk

Abstract

Vocal Tract Length Normalisation (VTLN) is a commonly used technique to normalise for inter-speaker variability. It is based on the speaker-specific warping of the frequency axis, parameterised by a scalar warp factor. This factor is typically estimated using maximum likelihood. We discuss how VTLN may be applied to multiparty conversations, reporting a substantial decrease in word error rate in experiments using the ICSI meetings corpus. We investigate the behaviour of the VTLN warping factor and show that a stable estimate is not obtained. Instead it appears to be influenced by the context of the meeting, in particular the current conversational partner. These results are consistent with predictions made by the psycholinguistic interactive alignment account of dialogue, when applied at the acoustic and phonological levels.

1. Introduction

It is well known that the speech signal carries information about vocal tract length (VTL): for example, the formant frequencies of vowels decrease as the VTL increases [1]. VTL normalisation (VTLN) is now a commonly used normalisation technique in speech recognition [2, 3, 4, 5, 6], that involves a speaker-specific (or speaker cluster-specific) warping of the frequency axis. The speaker-specific warp factor is usually obtained by maximising the likelihood with respect to the model.

Maximum likelihood (ML) estimation of VTLN warp factors only indirectly normalises the spectrum to account for VTL: there are other factors (such as systematic pronunciation variation) which may also be normalised by spectral warping. Irino and Patterson [7] have suggested that VTL information can be extracted directly, and have proposed an auditory-inspired transform which separates VTL size from shape information. This account has been supported by some recent perceptual experiments [8], which provide evidence for the hypothesis that the auditory system automatically normalises for VTL when processing speech or other vocalised sounds.

In this paper we are concerned with applying VTLN to multiparty conversations in a meeting environment. Most successful applications of VTLN have been reported for conversational telephone speech tasks, where there are distinct speaker sides and usually several minutes of speech per speaker. However in the case of meetings, even if speaker segmentation is available, the amount of speech data per speaker can vary significantly, making it difficult to obtain stable estimates of the VTLN warping factor. We have performed experiments using the ICSI meetings corpus [9], using the NIST Spring 2004 Meeting Evaluation data for development and test [10].

In addition to reporting VTLN results on this data we have investigated the stability of the estimated speaker-specific warping factors. Although the length of a speaker's vocal tract is dependent on the positions of the lips and the larynx, to a first approximation it may be regarded as constant. The fact that the speaker specific warping factors estimated by ML VTLN vary over time indicates that the frequency warping estimates are compensating for more than just VTL. In particular, we have investigated the relationship between the frequency warping factor and the addressee of the current speaker.

2. VTLN

VTL has a substantial effect on the observed spectrum: for example, a typical female speaker exhibits formant frequencies around 20% higher than those of a male speaker. Cohen et al [2] reported that a linear warping of the frequency axis could compensate for such difference in VTL, resulting in a speech recognition system with a reduced word error rate (WER). Over the past 10 years VTLN has become a standard normalisation technique in speaker independent speech recognition, proving particularly effective in the domain of conversational telephone speech (CTS) [3, 5, 6]. Different warping techniques have been reported in the literature: frequency warping both linear [5] and exponential non-linear [11] and Bark/Mel scale warping [12]. Here, the frequency axis is warped with a factor α per speaker, by scaling the centre frequencies of the mel filterbank prior to the extraction of cepstral features.

Two main methods have been developed to compute α: ML, and a parametric approach. In the ML approach, the warping factor is estimated in order to maximise the probability of recognising an utterance given a particular acoustic model [3, 5, 6], whereas the second method derives the warping factor from estimated formant positions [11, 4]. Although the ML approach is computationally expensive, it is robust and consistent with the overall optimisation of the speech recogniser, since it maximises the likelihood—something not guaranteed by the second approach. Furthermore, the estimation of formant positions requires voiced segments only and this can be challenging with conversational natural speech [12] because it requires an accurate alignment, whereas ML does not have the same requirement.

Estimating α by ML increases the matching score with the acoustic models, thus making the warping factor very model dependent. Moreover, the estimated warping factor is stable only when a considerable amount of data is available. This is well matched to tasks such as CTS where homogeneous speaker sides are available for every speaker, but it is an issue to be addressed for domains such as meetings or broadcast news [13], where the amount of data per speaker varies consistently.

We have adopted an ML approach, employing a piecewise linear frequency warping with lower and upper cutoff frequencies [5, 14]. The warping factor α is estimated, using a Brent search based on quadratic interpolation, since the log-likelihood's trend of a given transcription tends to have a parabolic shape in function of the warping factor value. Then to maximise the likelihood of the normalised acoustic observation X^{α} given a transcription W and an acoustic model λ the following equation has to be solved:

$$\alpha = \arg\max_{\alpha}(Pr(X^{\alpha}|\lambda, W)) \quad (1)$$

We applied VTLN both during training and testing. For training we adopted an iterative procedure with the following steps [5]:

1. warping factors estimation using a non-normalised model and normalised feature computation using those warping factors
2. training pass: single-pass retraining [14] starting from non-normalised models and a few Baum-Welch passes
3. warping factor estimation using the previous pass acoustic models and normalised features computation
4. training pass: like step 2 but starting from normalised models of the previous pass
5. repeat steps 3 and 4 until WER on the development data set stabilises

This allows warping factors to converge, providing αs in the range between 0.8 and 1.2, with the distribution of warping factors for female speakers decreasing to less than 1, and the distribution for males increasing to greater than 1.

For testing we adopted a 2 pass procedure [6, 5]:

1. decode using non-normalised features and models
2. evaluate warping factors using normalised models and the preliminary transcription of previous pass
3. normalise acoustic features and decode using normalised models

3. Speech recognition experiments

The experiments we report in this paper have been performed on the ICSI meeting corpus [9]. This is a collection of 75 multiparty meetings of research groups (approximately 72 hours in total) with an average of 6 participants per meeting and 53 total participants. The meetings consist of unconstrained natural and spontaneous speech. Many of the speakers are involved in several different meetings in the corpus. We used 70 of these 75 meetings as training data. For testing we used the ICSI portions of the NIST Spring 2004 Meetings Evaluation development and evaluation sets (referred to as RT04sdev and RT04seval, respectively) [10]. Each of these test sets contains 10 minutes of 2 different meetings, with 12 different speakers in RT04sdev and 15 in RT04seval.

We obtained baseline acoustic models for this corpus using a training set consisting of 300 hours of CTS from the Switchboard and Callhome corpora (referred to as h5train03 [15]). The resultant models (cross-word triphones trained on conversational side based cepstral mean normalised PLP features) were then MAP adapted to the meeting domain using 70 of the 75 ICSI meetings. VTLN training was performed, starting from these MAP adapted models, using an iterative procedure as described above. Each intermediate model was tested on both test sets (using a bigram language model and a vocabulary of 50k words), and the results are shown in table 3. [1] Moreover Cepstral Mean Normalisation (CMN) and Cepstral Variance Normalisation (CVN) have been adopted both during training and testing where mean and variance have been calculated over a complete channel for every speaker per meeting [5]. Only two VTLN training passes were required for convergence of the distribution of warping factors, although after convergence some small ripples in the WER may be observed.

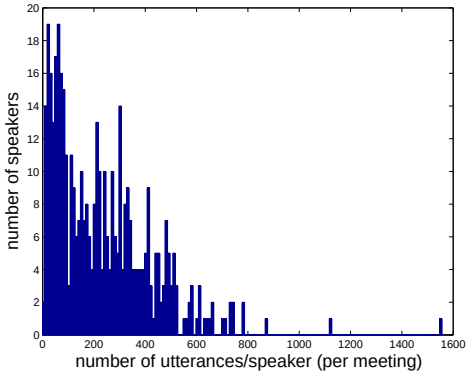

Figure 1: Distribution of the number of utterances per speaker (per meeting) for the ICSI training dataset

	RT04sdev	RT04seval
noVTLN	27.0	34.2
VTLN 1	24.6	31.6
VTLN 2	24.5	31.2
VTLN 3	24.9	32.1
VTLN 4	24.4	31.3
VTLN 5	24.3	31.0

Table 1: Speech recognition results of VTLN experiments in % WER for five successive training passes of the iterative procedure.

4. Warping factor estimation

The amount of data per speaker in each meeting varies considerably with a minimum of 3 seconds to a maximum of more than 1 hour of speech per speaker per meeting with an average utterance duration of about 2.4 seconds in the training set. This feature of the meeting data affects the reliability of the VTLN warping factor estimates. Fig. 1 shows the distribution of the number of utterances per speaker—about 33% of the speakers have less than a hundred utterances per meeting.

Fig. 2 illustrates how the estimated warping factor depends on the number of utterances from which it is estimated. This behaviour is seen for most speakers. Here CMN and CVN have also been performed using different amounts of data. The ML estimate for the VTLN warping factor takes around twenty utterances before it begins to stabilise. If the estimated warping factors do indeed correspond to normalising for variability in VTL between speakers, then we would expect their estimates to

[1] Different warping factors were estimated for those speakers that occurred in both sets.

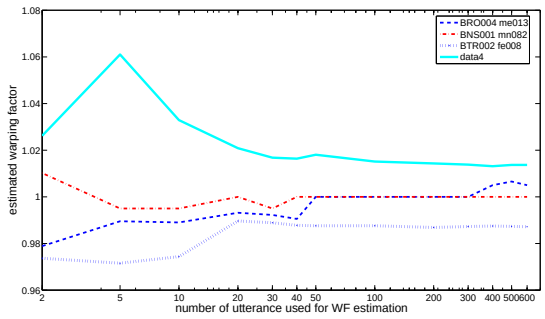

Figure 2: Trend of the warping factor values using different amount of utterances for the estimation

be more stable. This variability is highlighted if we compute the warping factor as a moving average across ten utterances (fig. 3).

Multiparty meetings are characterised by a rich speaker turn structure, and we have investigated the influence of this on the warping factor estimates. In particular, we have investigated the dependence of the warping factor estimated for a speaker given the speaker that they are addressing. Accurate labelling of which participant(s) each utterance is addressed to is rather labour intensive—and can be difficult from an audio-only recording of a meeting. We have made the approximation that a speaker speaking at a given time is addressing to the most recent speaker (not including backchannel-type utterances).

For each utterance of each speaker we estimated a local warping factor using that utterance and the previous nine utterances. Our first question was whether the distribution of the warping factor for speaker A ($wf(A)$) has a dependence on the previous speaker. We used a hypothesis testing procedure to do this, where the null hypothesis H_0 is that the mean value of the warping factor of speaker A given that s/he spoke after speaker B is equal to the global warping factor value for A computed using all the data for that meeting. The probability to accept H_0 has been computed as $P(t)$ with:

$$t = \frac{wf(A) - \mu(wf(A|B)}{\frac{\sigma(wf(A|B))}{\sqrt{n}}} \qquad (2)$$

where $\mu(wf(A|B))$ is the mean warping factor of A after B, σ is the standard deviation and n is the number of data (utterances) considered.

We studied eight meetings from the ICSI training dataset taken from different meeting types [9] and in a way that some of the speakers were present in more than one meeting. Using the Student t-test ($p = 0.05$) we found that for 84% of the speaker pairs the mean warping factor $\mu(wf(A|B))$ was significantly different from the global warping factor for A. Thus it appears that the turn taking process has some influence over warping factors. We also performed an unpaired t-test on the distributions of the warping factors of $A|i$ and $A|j$ for every speaker $i \neq A$ and $j \neq A$ with $i \neq j$. Here the null hypothesis H_0 is that the mean warping factor of $A|i$ and $A|j$ is the same. At 5% significance we found that in 78% of the cases the means of the two distributions were significantly different and we could reject the null hypothesis. Therefore it is likely that a given speaker A will speak differently according to whom they are addressing and that the ML estimate of the warping factor takes this into account.

We performed a preliminary speech recognition experiment computing for every speaker a different warping factor for every possible speaker turn. We tested on a set of 5 complete meetings from the ICSI corpus (referred to as amieval [15]) which were excluded from the training. We compared normalising with a global warping factor per speaker with normalising with warping factors conditioned on the previous speaker. Initial results indicated that the WER obtained without VTLN (32.6%) was significantly improved by both global speaker warping (27.1%) and speaker-conditioned warping (28.0%), but (in this initial experiment) no improvement was found using speaker-conditioned warp factors. Work in progress is using a moving average estimation of VTLN warp factors.

5. Interactive alignment

Fig. 3 (bottom) plots $wf(i|j)$ and $wf(j|i)$ against time. It shows the local warping factor estimated for speaker *me003* for utterances following utterances by speaker *me012* and vice versa (*me012* after *me003*) for the *BED003* meeting. This figure may be segmented in a sequence of intervals: segments where the 2 warping factor sequences show a similar behaviour (aligned) and segments where the warping factor dynamics are nonaligned. Typically 25–30% of the segments are aligned. A similar structure can be also observed for the fundamental frequency F0 (fig. 3, top) which plots the mean F0 value for each utterance.

This structure corresponds well to a psycholinguistic account of dialogue, referred to as the *interactive alignment model* [16]. In this account of dialogue it is argued that linguistic comprehension and production representations are shared between interlocutors in a dialogue. This is referred to as *alignment* and it is argued that it occurs at many levels: phonetic, phonological, lexical, syntactic and semantic. Interactive alignment is manifested at these different levels within a dialogue, for example the use of similar syntactic structures, lexical repetitions, and common pronunciations. Krauss and Pardo [17] have suggested that alignment in dialogue may be clearly observed at the phonological level and have presented preliminary evidence based on the vowel space (in terms of the first two formants) of interlocutors in two party dialogues. Their results suggest that the parties in a dialogue align at the phonological level as initially divergent pronunciations converge as the dialogue progresses. Kakita [18] has presented evidence of the convergence of F0 between parties in a dialogue.

The behaviour of the warping factor estimates is in line with the interactive alignment account of dialogue. The estimated warping factors of two interlocutors are typically non-aligned at the start of a meeting, but can be seen to align (or at least go through phases of alignment) as the meeting progresses. In addition to the length of the vocal tract, there is a well known relationship between the VTLN warping factor and F0 [11, 4]. However, the alignment between speakers' warping factors is not entirely accounted for by F0 and a shift in formant frequencies caused by interactive alignment at the phonological level is also being captured by the frequency warping factors estimated by the VTLN procedure.

6. Conclusions

This paper has two main contributions. Firstly we have demonstrated that ML VTLN may be applied to speech recognition of meeting recordings resulting in a relative decrease in WER of over 15%. Secondly we have demonstrated that the frequency

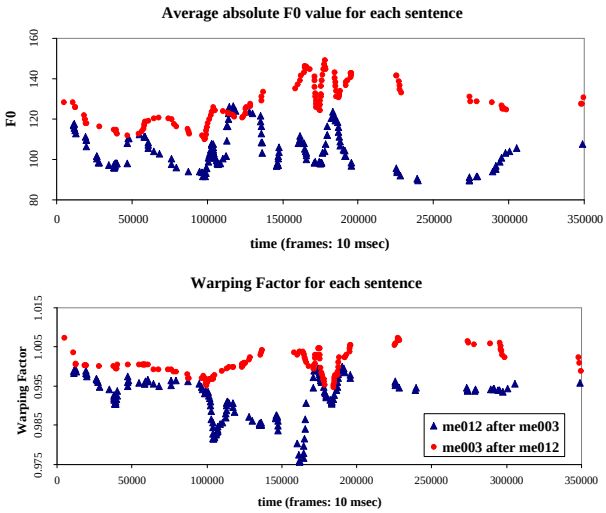

Figure 3: Trend of the warping factor of 2 speakers: me012 after me003 and me003 after me012

warping factors estimated within VTLN are not simply warping the spectrum to take account of interspeaker variability in vocal tract length. Our results indicate that the VTLN warping factor estimated for a speaker co-varies with the warping factor estimated for the current conversational partner, and that these coordinated variations result from alignment of F0 and from phonological alignment.

These results have implications for acoustic modelling of multiparty conversations and suggest some promising directions for future research:

1. The development of acoustic models of multiparty speech with a dependence on the other conversational participants. This would imply that it is more consistent at the acoustic modelling level to perform recognition of complete meetings rather than meeting segments.
2. The development of conditional pronunciation models that take advantage of phonological alignment between conversational parties.
3. Investigation of direct techniques for the separation of vocal tract size from shape (eg [7]), rather the indirect methods currently employed.

7. Acknowledgements

This work was partly supported by the EU 6th framework IST Integrated Project AMI (Augmented Multi-party Interaction) FP6-506811 (ref: AMI-68). Thanks to the University of Cambridge Engineering Department Speech Group for the use of h5train03 and HDecode. And last but not least we'd like to thank everybody in the AMI ASR work group for their work, support and advices [15].

8. References

[1] G. Fant, *Acoustic Theory of Speech Production*. The Hague: Mouton, 1960.

[2] J. Cohen, T. Kamm, and A. Andreou, "Vocal tract normalization in speech recognition: compensating for systematic speaker variability," *J. Acoust. Soc. Am.*, vol. 97, no. 5, Pt. 2, pp. 3246–3247, 1995.

[3] L. Lee and C. Rose, "Speaker normalisation using effi cient frequency warping procedures," *Proc. IEEE ICASSP*, pp. 353–356, 1996.

[4] S. Wegmann, D. McAllaster, J. Orloff, and B. Peskin, "Speaker normalisation on conversational telephone speech," *Proc. IEEE ICASSP*, 1996.

[5] T. Hain, P. Woodland, T. Niesler, and E. Whittaker, "The 1998 HTK system for transcription of conversational telephone speech," *Proc. IEEE ICASSP*, 1999.

[6] L. Welling, H. Ney, and S. Kanthak, "Speaker adaptive modeling by vocal tract normalisation," *IEEE Transactions on Speech and Audio Processing*, vol. 10, pp. 415–426, Sept. 2002.

[7] T. Irino and R. Patterson, "Segregating information about the size and shape of the vocal tract using a time-domain auditory model: the stabilized wavelet-Mellin transform," *Speech Communication*, vol. 36, pp. 181–203, 2002.

[8] D. R. R. Smith, R. D. Patterson, R. Turner, H. Kawahara, and T. Irino, "The processing and perception of size information in speech sounds," *J. Acoust. Soc. Am.*, vol. 117, no. 1, pp. 305–318, 2005.

[9] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters, "The ICSI meeting corpus," *Proc. IEEE ICASSP*, 2003.

[10] "Rich transcription 2004 spring meeting recognition evaluation website," *http://www.nist.gov/speech/tests/rt/rt2004/spring/index.htm*, 2004.

[11] E. Eide and H. Gish, "A parametric approach to vocal tract length normalization," *Proc. IEEE ICASSP*, pp. 346–348, 1996.

[12] P. Zhan and A. Waibel, "Vocal tract length normalization for large vocabulary continuous speech recognition," *CMU Language Technologies Institute Technical Report*, May 1997.

[13] D. Kim, M. Gales, T. Hain, and P. Woodland, "Using VTLN for broadcast news transcription," *ICSLP*, 2004.

[14] S. Y. et al., "The htk book," *Revised for HTK Version 3.2*, December 2002.

[15] T. Hain, J. Dines, G. Garau, M. Karafi at, D. Moore, V. Wan, R. Ordelman, I.Mc.Cowan, J.Vepa, and S.Renals, "An investigation into transcription of conference room meetings," *Submitted to Eurospeech*, 2005.

[16] M. J. Pickering and S. Garrod, "Towards a mechanistic psychology of dialogue," *Behavioural and Brain Sciences*, vol. 27, pp. 169–226, 2004.

[17] R. M. Krauss and J. S. Pardo, "Speaker perception and social behaviour: Bridging social psychology and science," 2004, http://www.columbia.edu/∼rmk7/PDF/Bridges.pdf.

[18] K. Kakita, "Inter-speaker interaction of F0 in dialogs," *Proc. ICSLP*, 1996.