INTERSPEECH 2009 BRIGHTON

Speaker normalization for template based speech recognition

Sébastien Demange, Dirk Van Compernolle

Katholieke Universiteit Leuven - Dept. ESAT
Kasteelpark Arenberg 10, B-3001 Leuven, BELGIUM
{sebastien.demange,dirk.vancompernolle}@esat.kuleuven.be

Abstract

Vocal Tract Length Normalization (VTLN) has been shown to be an efficient speaker normalization tool for HMM based systems. In this paper we show that it is equally efficient for a template based recognition system. Template based systems, while promising, have as potential drawback that templates maintain all non phonetic details apart from the essential phonemic properties; i.e. they retain information on speaker and acoustic recording circumstances. This may lead to a very inefficient usage of the database. We show that after VTLN significantly more speakers - also from opposite gender - contribute templates to the matching sequence compared to the non-normalized case. In experiments on the Wall Street Journal database this leads to a relative word error rate reduction of 10%.

Index Terms: template based speech recognition, speaker normalization, VTLN

10.21437/Interspeech.2009-200

1. Introduction

Template based speech recognition [1, 2] has recently aroused a revival of interest. It recognizes speech by comparing any input signal to templates (also referred to examples or episodes) stored in a database. Contrary to hidden Markov models (HMMs), episodic modeling preserves the trajectories and the durations of past experienced episodes. In addition, acoustic details contained in a template database are available and can be used during the decoding. However, the very high variability of speech results in very sparse distributions of the templates in the acoustic space and a strong overlap between examples of different phones. This recognition paradigm is then very exposed to non-linguistic speech variability which is considered as noise in speech recognition.

The shape of the vocal apparatus, and more specifically the length of the vocal tract, is one of the major sources of inter-speaker variability. Usually the vocal tract length (VTL) of women is approximately 15% shorter than the VTL of men causing the formant frequencies of women to be higher. Vocal tract length normalization (VTLN) is a widely used method for reducing this inter-speaker variability and has been demonstrated efficient for HMM based speech recognition. VTLN rescales, according to a warping factor (WF), the frequency axis of the acoustic feature vectors so that the observations are more similar across different speakers.

We propose in this work to investigate the compatibility and the usefulness of VTLN in the context of exemplar based speech recognition. Indeed, one can question the need for normalizing the speech signal. Firstly, the use of any signal normalization mechanism is conflicting with one of the episodic modeling advantages as normalizing the signal could have negative effects on the acoustic trajectories. Secondly, while an input speech signal can equally target HMM distributions relying mainly on male or female training samples, template based recognition ensures a full speaker consistency when matching a particular example to a part of the signal to be recognized. This consistency can even be extended to the whole sentence by means of little extra processing. Since, a template database is usually made of tens of hours of records collected from a large number of speakers, we can assume it contains at least one voice close enough to the test voice to provide the recognition engine with good examples. On the other hand, the VTL does not explain all the speech variability. Many other factors can influence the realization of a particular phone such as its phonetic contexts. Therefore, only a very small fraction of the template database can contribute valuable examples because useful examples shoud have been produced in similar conditions as the input signal. Applying VTLN on both the template database and the input signal would weaken the speaker consistency requirement. Then, much more speakers present in the database would be able to furnish good candidates and this would result in more possibilities to find a good template targeting the other conditions. Moreover, reducing the inter-speaker variability should also improve the distance between a template and the input signal since this distance would more rely on pure acoustic features than on speaker-dependent characteristics. As a consequence, the VTLN should also improve the recognition accuracy.

This work answers two questions: 1) Does VTLN improve the usage of the template database during the template selection? 2) Does VTLN offer the same recognition improvement observed with the HMMs and why? This paper is organized as follow. First a brief description of the template based speech decoder is given. Then, we present the VTLN technique in section 3. The experiment setup is given in section 4 and the experimental results are presented in section 5.

2. Template based speech recognizer: overview

Our template based speech recognizer relies on a two-layered decoder architecture.

First, the input signal is processed by a HMM based speech decoder. It results in a dense phone graph containing the more likely recognition hypotheses. Then, for each phone arc within the phone graph we select from the database the N-best matching templates. The timestamps as well as the phone identity of the phone arcs constrain the search. The distance between the input speech segments (aligned with the phone arcs) and each individual template is computed by Dynamic Time Warping (DTW) using locally scaled distances. A template graph is thus constructed. A template arc is defined by a start time, an end time, a template ID, a phone ID, and a score reflecting how close the input signal and the template are.

Secondly, the template graph is enriched with concatenation cost arcs which are inserted between all consecutive template arcs. They express, through a cost value, how smooth the concatenation of two templates is. Predefined additive costs are introduced each time the acoustic contexts of a template in the template graph differ from its original acoustic contexts and each time two templates coming from speakers of opposite gender are concatenated. Finally a cost is introduced for all template concatenations except if the templates were "natural" successors in the database. The template graph is then decoded using a finite state transducer combining the lexicon and the language model. The concatenation costs are defined locally but their influence is global. This way, the recognition result corresponds to a coherent template sequence which matches well enough the input signal.

3. Vocal tract length normalization

Several methods for estimating the WFs have been proposed. The most popular rely on an exhaustive search over all possible WFs to maximize the likelihood of the signal [3]. The general principle is to adapt the input signal to the acoustic models. Such an approach is not feasible with template based speech recognition since by definition it doesn't rely on acoustic models but on a huge collection of templates with many different speakers and it is thus impossible to adapt the input signal to all templates. We have chosen a fast on-line WF estimation proposed by Duchateau and al. [4] suitable for this approach. The WF estimation is done on a sentence by sentence basis, so that each individual template is rescaled with a unique WF. This way the normalization should preserve the acoustic trajectory.

We first define two classes of speakers (male and female) and assign a target WF to each of them: WF_{Male} and WF_{Female}. These target WFs represent the extreme values which can be used. They can be seen as the WFs for the lowest male voice and the highest female voice. For each input signal X, the WF is estimated as a weighted sum of the male and female target WFs.

$$WF = w_{Male}.WF_{Male} + w_{Female}.WF_{Female} \quad (1)$$

The weight w_{Male} is the probability that the current speaker is a male, computed from two Gaussian mixture models $\mathcal{M}_{Male}$ and $\mathcal{M}_{Female}$ of male and female generic speech (w_{Female} is obtained similarly).

$$w_{Male} = \frac{P(X|\mathcal{M}_{Male})}{P(X|\mathcal{M}_{Male}) + P(X|\mathcal{M}_{Female})} \quad (2)$$

The probability $P(X|\mathcal{M}_{Male})$ that the current speaker is a male is expressed as the geometric mean of each speech frame likelihood raised to the power of D:

$$P(X|\mathcal{M}_{Male}) \sim \left[\prod_{i=1}^{n} P(x_i|\mathcal{M}_{Male})\right]^{\frac{D}{n}} \quad (3)$$

The constant D is a parameter which controls the distribution of the estimated WFs. Setting D to a high value makes most of the estimated WFs close to one of the targets while using a smaller value softens the decision shifting the WFs toward 1. Figure 1 shows the distributions of the WFs estimated on the template database with $D = 0.1, 0.5, 1.0$ and 2.0 and with WF_{Male} and WF_{Female} respectively equal to 0.85 and 1.15. Our better recognition accuracy with this WF estimation was obtained with $D = 0.5$. This value leads to two Gaussian like distributions of the WFs similar to what we can expect for a human VTL distribution. Moreover, a comparison with a classical grid search approach (Table 1) shows that this WF estimation is efficient in complexity and performance. Finally, experiments over different corpora (not shown here) suggest that the optimal parameters generalize well across language and database.

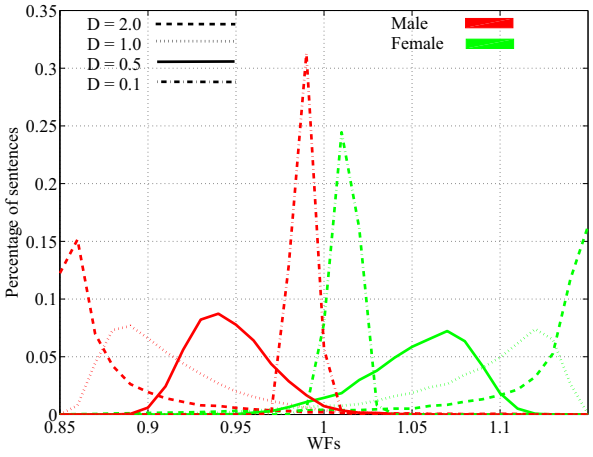

Figure 1: Distribution of estimated WFs on WSJ0+1 training corpus.

	Dev92		Nov92	
	WER	Rel. Imp.	WER	Rel. Imp.
No VTLN	3.76	-	2.54	-
Grid search	3.60	4.3	2.30	9.5
Our	3.30	12.2	2.32	8.7

Table 1: Comparison between our WFs estimation and a grid search approach on the Dev92 and Nov92 test sets of the Wall Street Journal corpus.

One can note that the frequency warping could also be processed during the template matching by searching the best frequency scaling together with the best time alignment. Published results have shown the effectiveness of such an approach [5]. However, the increase of the computational load is excessive for the template matching approach.

4. Experimental set-up

4.1. Corpus

The results presented in the next section are obtained using the WSJ (Wall Street Journal) continuous speech recognition corpus. The template database is built from the WSJ0 and WSJ1 (WSJ0+1) training sets (284 speakers) and consists of 2 800 000 single-phone templates. The database phone segmentation is obtained by forced alignment using our in house HMM system. The 15 845 cross-word triphones are trained on the WSJ0+1 training set. The templates are stored in the order they appear in the corpus, so that it is possible to access their original phonetic context. For the evaluation, only the 5k close vocabulary development (Dev92) and evaluation (Nov92) sets are used.

4.2. Feature extraction

The front-end consists of a 36-dimensional feature vector each 10 ms of speech using overlapping frames of 25 ms. First, 24 Mel-scaled filterbank coefficients are extracted. Secondly, VTLN is applied and the feature vector is extended by the first and second time derivatives. Finally, MIDA, an improved LDA algorithm, is used to transform the 72 dimensional space and keep the 36 most informative directions.

4.3. Template graph and concatenation costs

The phone graphs are deduced from word graphs produced by the HMM based speech recognizer preventing any dead-end when decoding the template graphs. The template graphs are built from the phone graphs by selecting the 50 best matching templates for each phone arc. The concatenation costs as well as the language model weight are optimized on the development set using a genetic algorithm. For each experiment the optimization was stopped after approximately 250 iterations since no improvement was observed over the last 50.

5. Evaluation

The effect of the VTLN on the template based decoder is presented in this section. We first evaluate how it effects the phone graphs. Second, we assess its impact on the template selection and finally we present the recognition results.

5.1. Phone graph generation

Table 2 presents the HMM based recognition results obtained with and without VTLN on the Dev92 and Nov92 sets as well as some statistics of the resulting phone graphs. The first row shows the WER when considering the best path (maximum HMM likelihood) in the phone graph. The second row shows the best WER which can be achieved when considering all paths within the graph. The third, fourth and fifth rows describe the phone graphs w.r.t. to the *density* (average number of context independent phones in parallel per frame), the *event rate* (average number of context dependent phones that start per frame) and the *fan out* (average number of arcs leaving a node).

	Dev92		Nov92	
	novtln	vtln	novtln	vtln
Recognition result				
WER (%)	3.76	3.30	2.54	2.32
Phone graph description				
Min WER (%)	0.46	0.41	0.26	0.26
Density	1.49	1.51	1.42	1.45
Event rate	0.66	0.68	0.55	0.60
Fan out	1.62	1.64	1.63	1.67

Table 2: Comparison of the phone graphs generated with and without VTLN.

We can observe relative improvements of 13% and 8.5% on the WER respectively on the Dev92 and Nov92 sets when using VTLN. It indicates that VTLN contributes to increase the likelihood of the correct path compared to the other paths in the graph. However, the different graphs still have similar properties and more important it is potentially possible to achieve the same recognition performance from these graphs independently of the use of VTLN. Because our decoder in this configuration doesn't use the HMM likelihoods we can thus expect a very small impact of the VTLN on the phone graphs. Nevertheless, VTLN can improve the time alignment of the phone arcs within the graph, i.e. the phone boundaries can better fit the input signal. Such effect cannot be properly measured but can improve the matching and consequently the WER. This is why we propose in section 5.3 to assess the impact of VTLN in terms of WER improvement over four experiments with combined use of VTLN for the phone graph generation and the template selection.

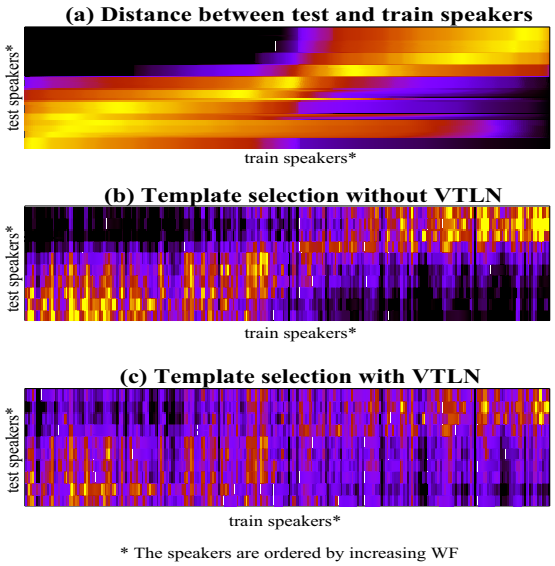

Figure 2: Visual assessment of the origin of the best matching templates: (a) distances between test and train speakers, (b) and (c) percentage of templates selected from train speakers for each test speaker.

5.2. Template selection

Figure 2 gives a visual assessment of the template selection over the training speakers. For each panel, the vertical and horizontal axis represent respectively the 10 and 284 speakers of the development and training sets. The speakers are ordered by increasing WFs. The top panel shows the distance between the test and train speakers. We have chosen to express the distance between two speakers as the difference between their average WF (so implicitly by the difference of their VTL).

$$Dist(spkr_1, spkr_2) = \left| WF_{spkr_1} - WF_{spkr_2} \right| \quad (4)$$

The brighter the color is, the closer the speakers are. Four distinct parts emerge from this panel. The bottom-left and top-right quarters represent the distances between speakers of same gender while the top-left and bottom-right quarters represent distances between speakers of opposite gender.
The middle and bottom panels give for each test speaker the percentage of templates selected from each train speaker. Here, the darker the color is, the less a particular train speaker contributes templates for a particular test speaker. It is clear from the middle panel that the inter-speaker variability contributes to under-utilization of the database. Indeed, the similarity between the top and middle panels clearly indicates that most of the templates for a particular test speaker come from a train speaker of

the same gender with in addition a preference for speakers with similar VTL. After the speaker normalization (bottom panel) the templates are selected from more different speakers even if they strongly differ from the test speakers and/or if they are of opposite gender.

Table 3 provides detailed statistics about the template selection and shows that VTLN affects more the extreme test speakers. As a result the coverage of the database during the template selection is increased and is similar over all test speakers. The percentage of selected templates which match the gender of the test speaker drops from 75% to 60% while the number of different speakers from whom the selected templates come from increases from 210 to 230 ($\approx +10\%$). The last two columns reveal that the number of speakers of opposite gender is on average increased by 8% (absolute) indicating that most of the new speakers from whom the templates have been selected are of opposite gender than the test speakers. Thus, the speaker normalization allows a more efficient usage of the database.

Dev spkr		% templates matching the gender		# train spkrs per test spkr		% train spkrs matching the gender	
ID	WF	no vtln	vtln	no vtln	vtln	no vtln	vtln
053	1.082	79%	59%	203	231	64%	54%
420	1.079	84%	67%	192	223	69%	57%
050	1.068	83%	63%	193	224	68%	57%
421	1.039	68%	58%	223	235	58%	52%
422	0.986	43%	42%	231	233	47%	47%
423	0.962	70%	60%	222	234	59%	54%
052	0.937	72%	59%	223	231	59%	55%
051	0.927	80%	61%	212	231	62%	54%
22g	0.924	83%	69%	206	229	67%	58%
22h	0.917	83%	62%	202	233	69%	56%
All		75%	60%	210	230	62%	54%

Table 3: Evaluation of the effect of VTLN during the template selection on the development set (line 1-5: female speakers, line 6-10: male speakers).

5.3. Word recognition

Four recognition experiments have been done with combined use of VTLN for both the phone graph generation and the template selection. The results are summarized in table 4.
Using VTLN for the phone graph generation gives only a non-significant 2% relative improvement on both the Dev92 and Nov92 test sets while gains of 5% and 9% are obtained when applying VTLN for the template selection. As expected the impact of VTLN is smaller on the phone graph than on the template selection. A deeper examination of the recognition results reveals that the percentage of templates of opposite gender being part of the best template sequences increases from 9% to 15% clearly indicating that VTLN not only improves the matching with templates of opposite gender but also increases their contribution to the winning template sequence. An attempt was done without using the gender mismatch concatenation cost. While the proportion of templates of opposite gender increased from 15% to 32%, the WER remained stable.

The experimantal results corroborates our initial intuition. Indeed, VTLN allows to release the gender consistency requirement. As a consequence much more examples are available for satisfying the phonetic context consistency. The new templates are very valuable as they significantly contribute to the classification decision leading to a 10% relative improvement in the WER.

Phone graph	no vtln		vtln	
Template selection	no vtln	vtln	no vtln	vtln
Dev92	5.37	5.16	5.29	5.00
Nov92	4.86	4.30	4.67	4.33

Table 4: Comparison of WER with combined use of VTLN.

6. Conclusion

We have studied in this article how helpful VTLN is in the context of template based speech recognition. Furthermore, we have pinpointed the reason for the observed improvement. VTLN doesn't really affect the phone graph but it allows the system to make a better usage of the template database during the template selection. Indeed, many new templates become strong candidates for the recognition. Therefore, the number of potential good examples increases and the recognizer has more flexibility to satisfy the consistency of the winning template sequence w.r.t. the phonetic contexts. This results in a relative WER improvement of 10% on the WSJ database which is equivalent to the improvement obtained with a classical HMM based speech recognizer.

7. acknowledgment

This work is supported by the Sound-to-Sense project funded by the EU Marie Curie Research Training Network (MC-RTN) and the FWO project G.0260.07 "TELEX" funded by the Flemish Research Foundation.

8. References

[1] Mathias De Wachter, Mike Matton, Kris Demuynck, Patrick Wambacq, Ronald Cools, and Dirk Van Compernolle, "Template based continuous speech recognition," *IEEE Trans. on ASLP*, vol. 15, pp. 1377–1390, May 2007.

[2] Viktoria Maier and Roger Moore, "Temporal episodic memory model: An evolution of minerva2," in *Proc. INTERSPEECH*, Antwerp, Belgium, Aug. 2007, pp. 866–869.

[3] Lutz Welling, Hermann Ney, and Stephan Kanthak, "Speaker adaptive modeling by vocal tract normalization," *IEEE Trans. on SAP*, vol. 10, pp. 415–426, September 2002.

[4] Jacques Duchateau, Mari Wigham, Kris Demuynck, and Hugo Van hamme, "A flexible recogniser architecture in a reading tutor for children," in *Proc. ITRW on Speech Recognition and Intrinsic Variation*, Toulouse, France, May 2006, pp. 59–64.

[5] Ming Liu, Xi Zhou, Mark Hasegawa-Johnson, Huang Thomas S., and Zhengyou Zhang, "Frequency domain correspondence for speaker normalization," in *Proc. INTERSPEECH*, Antwerp, Belgium, Aug. 2007, pp. 274–277.