Dynamic Vocal Tract Length Normalization in Speech Recognition

Daniel Elenius, Mats Blomberg
Department of Speech Music and Hearing, CSC, KTH, Stockholm

Abstract

A novel method to account for dynamic speaker characteristic properties in a speech recognition system is presented. The estimated trajectory of a property can be constrained to be constant or to have a limited rate-of-change within a phone or a sub-phone state. The constraints are implemented by extending each state in the trained Hidden Markov Model by a number of property-value-specific sub-states transformed from the original model. The connections in the transition matrix of the extended model define possible slopes of the trajectory. Constraints on its dynamic range during an utterance are implemented by decomposing the trajectory into a static and a dynamic component. Results are presented on vocal tract length normalization in connected-digit recognition of children's speech using models trained on male adult speech. The word error rate was reduced compared with the conventional utterance-specific warping factor by 10% relative.

Introduction

Mismatch between training and test conditions is a major cause of performance degradation in automatic speech recognition. Much effort has been invested into reducing this mismatch using adaptation and normalization techniques. A special adaptation category is based on predictive modelling (Gales, 1998). In this approach, explicit knowledge of how specific properties of the speaker or the acoustic environment affect the speech signal is applied to adapt the model to such conditions without the use of separate adaptation data. The technique has been used to compensate for mismatch in background noise (Gales, 1998), vocal tract length (VTL) (Lee and Rose, 1996) and voice source quality (Blomberg and Elenius 2009).

Adaptation based on predictive modelling typically uses the unknown test utterance itself as adaptation data. A transform, which operates on all frames of the utterance or all the trained models, is adjusted to maximize the likelihood of the decoding process. This procedure has been successfully applied in Vocal Tract Length Normalization (VTLN) on both adult and children's speech (Potamianos and Narayanan, 2003; Giuliani et al., 2006; Blomberg and Elenius, 2008). The transform in VTLN is a warping function which expands or compresses the frequency axis of the input signal or the trained model before matching.

However, there are arguments against the use of a time- and, accordingly, phoneme-invariant VTL transform. The effective vocal tract length of a speaker is dynamically increased/decreased by protrusion/spreading of the lips and by lowering/raising of the larynx (Fant, 1960; Dusan, 2007). Especially the larynx height can, to a large degree, be changed without shifting the perceived phonetic identity. This may give rise to intra- and inter-speaker variability for repeated pronunciations of the same word sequence. Another argument is that the length difference between two vocal tracts is in general not evenly distributed. For this reason, the frequency mapping function between their transfer functions is phoneme-specific. For example, phonemes with their main resonance frequencies belonging to the mouth cavity are expected to be quite insensitive to difference in pharynx length. The potential of phoneme-dependent warping has been demonstrated both in terms of formant frequencies (Fant, 1975) and cepstral deviation (Potamianos and Narayanan, 2003).

In recent years, increased interest has been directed towards a time-varying warp factor in VTLN to account for the above effects. Miguel et al. (2005) estimated a frame-specific warp factor by a three-dimensional Viterbi decoding process. Blomberg and Elenius (2007) searched for the best combination of warp factor specific phone models for the test

utterance. Maragakis and Potamianos (2008) used a two-pass method, where spectrally similar regions of the test utterance were transformed by the same warp factor. Elenius and Blomberg (2009) computed phoneme-specific warp factors for a group of children using adult models. Although they achieved systematic differences between the factor values, the use of these on another child group with the same age distribution improved the recognition accuracy only marginally.

The studies report little or moderate improvement from using phoneme-dependent and time-varying warp factors. However, the results should not be seen as a final assessment of the idea, since there are still approaches to the problem which have not yet been explored. In this paper, we propose and evaluate new methods for the implementation of dynamic warp factors. In short, a standard HMM is modified by extending each state by a number of warp factor specific sub-states. Constraints on the warp factor trajectory are implemented in the transition matrix. The trajectory can be specified to be constant or to have a constrained rate-of-change within realizations of phones or phone states. We have studied frame-wise adjustment in either of two ways: by unconstrained rate-of-change of the warp factor and by limiting the change to ±1 quantization step. Comparison is made with a phone-model-dependent warp factor, which gives an identical value for all instantiations in the utterance of a phone model, and with the conventional case, a time-invariant warp factor value for the whole utterance.

Although the speaker property used in this report is vocal tract length, the approach should be applicable also to other speaker and environment characteristics with time-varying behaviour, such as speech loudness, voice source quality, speech rate and fluctuating background noise.

Method

Dynamic modelling is accomplished by extending the HMMs of the acoustic model with a speaker characteristic dimension. New states are added in order to model property values deviating from those in the training data. The probability density functions of the new states are derived from the original state distribution by a speaker characteristic property transform. Thus, a probability density function for a new speaker characteristic property value is predicted based on the original pdf and a parametric transform. The rate-of-change constraints are implemented by means of changing the state transition probabilities in the extended model, which we will refer to as a Speaker Characteristic Augmented HMM (SCA-HMM). Further details of the method are given in (Elenius, 2010).

Phone model specific warping

Phone model specific warping does not require the SCA-HMM representation, and can be implemented as a standard HMM. The warp factors are estimated in a combined recognition-estimation step, in which a full recognition procedure is performed for each combination of factors considered. The search objective is to find the set of individually warped phone models, which maximizes the likelihood of the utterance. An exhaustive search is computationally very heavy and a reduction of the search space is required to make the search feasible. In this report, the warp factor is determined separately for each phone model, while the other models have an initial default warp factor value, as in Blomberg and Elenius (2007). Although the separate search has been shown to be sub-optimal, it is included here for comparison with other estimation methods.

Phone and sub-phone instance specific warping

The spoken realization of a phoneme is influenced by several variability sources, such as phonetic context, its position in the utterance, within a stressed or unstressed word, etc. It is thus unlikely that two realizations in the same utterance would have identical warp factors. This reduces the efficiency of phone model specific values. To account for this effect, phone-instance-specific warp factors can be used.

There is also motivation for changing the warp factor within a phone, i.e., between sub-phone segments. One example is the occlusion and the release phase of unvoiced plosives, which have different characteristics and should be modelled differently. The occlusion phase is likely to consist mostly of background sounds which are speaker-independent, while the release has speaker characteristic features.

These segments need to be compensated differently.

Variability in the movements of the articulators may cause a need to change the warp factor on a frame-by-frame basis. In general, the rate-of-change is expected to be slow due to the limited speed of the articulators, but more rapid changes might be required at phoneme boundaries and in transitional regions. To model this effect, different constraints should be used within and between phone-instantiations. We have chosen unconstrained rate-of-change at transitions between models. This is also motivated from a computational point of view, since constraining the trajectory across model boundaries would require an extensive increase in complexity using the current approach.

As mentioned above, the expansion of a standard HMM into an SCA-HMM adds sub-states to each original HMM state, where each of the new states represents a transformed version of its source state. Transitions between states are then added to model the different types of factor dynamics. The structure of the new SCA-HMM is indicated in Figure 1. We will refer to the states of the original HMM as main-states in the SCA-HMM. The new warp-specific states will be described as sub-states of the original main-states.

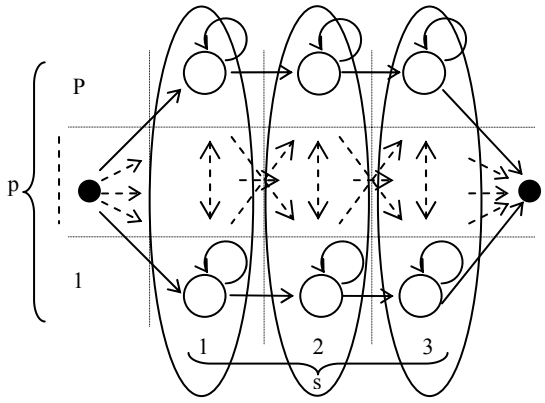

Figure 1. An SCA-HMM with a speaker-property value index, p, mixture, m, and main-state index, s. The original model is a three-state left-to-right HMM.

Constraints on the rate-of-change are implemented by selecting a subset of the possible transitions. In cases of more than one connection from a sub-state to other sub-states, uniform transition probability distribution is used.

Phone-instance-specific warping is realized by only allowing transitions between sub-states with the same warp factor. This will force the warp factor to remain fixed throughout the SCA-HMM. In Figure 1, this corresponds to having only horizontal connections between the main-states.

Sub-phone-specific warping is realized by adding transitions between sub-states of different warp factors between the main states of the phone-instance-specific model. In Figure 1, these are represented by non-horizontal connections between the main-states.

Frame-specific warp factors are implemented by adding intra-main-state transitions to the sub-phone-specific case. These correspond to vertical transitions within a main-state in Figure 1. Two types of possible transitions are used; one that enables a change to any other value and one that limits the factor change rate to one quantisation step. In the latter case, the same constraint is also applied to the inter-main-state transitions.

Static-dynamic factor decomposition

A straightforward candidate rule for constraining the allowed warp factor range during an utterance is to make certain that the range is sufficient for most speakers. However, this will permit the warp factor trajectory to vary between values corresponding to the shortest and the longest allowed vocal tracts in the same utterance. This is likely to limit recognition accuracy. We have approached this problem by decomposing the trajectory into a static, utterance specific, and a dynamic component, which are jointly optimized. The static component is estimated in a grid search while the dynamics, being modelled in the transition matrix, are determined in the Viterbi process. In this way, the dynamic range can be reduced by excluding inter-speaker variability of the static component. To reduce the increased computational demand of the joint two-fold search, existing speed-up techniques, such as a two-pass method (Lee and Rose, 1996) or a tree-based procedure (Blomberg and Elenius, 2009), can be applied.

Experiments

An experiment has been performed on the connected-digit corpus TIDIGITS. A speech recognition system was trained on the adult-

male part of the full training set. Evaluation was performed on the child part of the full TIDIGITS test set. The child subset consists of more than 3800 strings with more than 12600 digits.

In an experiment using a single component representation of the warp factor, the trained phone models were extended to SCA-HMMs by frequency warping with a warp factor in the range 1.0 to 1.7 with a step of 0.02. During recognition, warp factors were estimated specific to the utterance, a phone model, a phone instance, a phone state instance, or a speech frame. For the frame-wise estimation, an unconstrained rate-of-change of the warp factor was compared to one limited to ±0.02 between frames within a phone instance. In both cases, the inter-phone-instance change of warp factor was unconstrained.

Experiments with a static and dynamic component were performed using an SCA-HMM of 9 sub-states per main-state, spanning a dynamic interval of 0.16. Its centre represented the static component, which was determined by a grid search in steps of 0.02. The range of the combined warp method was limited to the same interval as above.

The experiments were performed using a connected-digit recognition system with triphone HMMs implemented in HTK. Each acoustic model consisted of 3 states with a GMM consisting of 16 mixtures and diagonal covariance matrices. A 39-dimensional acoustic feature vector was composed by 12 MFCCs and normalized log energy and their velocity and acceleration coefficients. Feature extraction was performed at a frame rate of 100 Hz with a 25 ms Hamming window and a mel-scaled filterbank of 38 filters in the range corresponding to 0 to 7.6 kHz.

Frequency warping was implemented as a piece-wise linear function using a linear transformation of models in the cepstral domain (Pitz and Ney, 2005). To avoid erroneous warping due to cepstral smoothing effects during transformation, training was performed using 18 cepstral coefficients. After warping the trained models, their mean and variance vectors were reduced to contain the 12 lower static, delta and acceleration coefficients, like in (Blomberg and Elenius, 2008).

Results and Discussion

The word error rates of the investigated methods are shown in Table 1. The original error rate using adult models was considerably decreased by all VTLN methods. Using single-component representation, none of the estimation units was significantly better than a standard utterance-specific warp factor. In contrast, the decomposed static and dynamic representation demonstrated clear superiority over the utterance-specific factor and all single-valued estimation units, with a minimum WER of 3.47% for state-instance-specific factor. Similar relations between the different estimation units were observed as for the non-decomposed case.

The superior result of the state-instance specific factors compared with that of the phone-instance condition supports the initial argument that the warp factor should be allowed to differ between sub-phonemic sound-events.

The error rate of the phone-model-specific factors is substantially higher than the other techniques. One reason for this might be found in the suboptimal search algorithm.

Table 1. WER of VTLN methods. The baseline result for original male models is 47.55%.

	Estimation technique	
VTLN estimation unit	Single component	Static & dynamic
Utterance	3.85	-
Phone model	6.47	-
Phone instance	4.19	3.69
State instance	3.84	3.47
Frame, Δ unconstrained	3.87	3.67
Frame, |Δ| <= 0.02	4.13	3.71

An example of chosen warp factor trajectories for a child's utterance (from the TIDIGIT training set) is shown in Figure 2. The unconstrained frame-specific warp factor exhibits a substantial variation from frame to frame. A high rate-of-change is used for acquiring the best match. Still, the repetitions of identical digits exhibit similar, regular, patterns.

The heavily rate-constrained frame-based trajectory is evidently too smooth to adjust to rapid intra-phone effects. It is probable that an intermediate rate limit would give a better

result. A few instances of abrupt change in the trajectory are explained by the fact that the rate was unconstrained at phone boundaries.

The sub-phone instance specific method results in a warp factor, which is constant within main-states and changes instantly at state transitions. It partly includes the abrupt changes observed in the frame-based method but removes within-state variation.

Static-dynamic decomposition can be observed to correct excessive warping of single component warping at several positions.

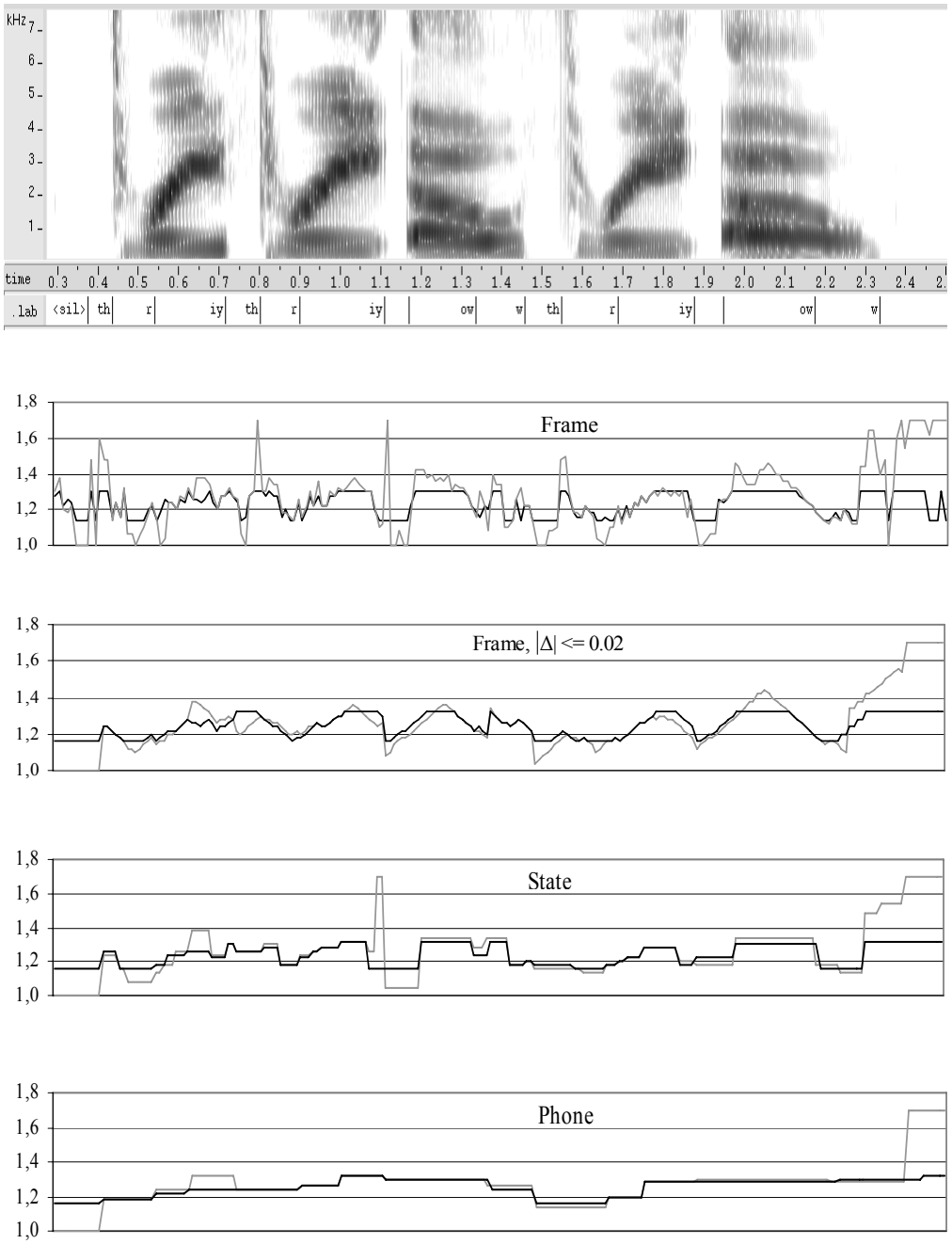

Figure 2. Warp-factor as a function of time for frame, state, and phone instance-specific warping for a boy's utterance "3 3 oh 3 oh". Gray and black curves represent the use of a single-component and a static-dynamic decomposed warp factor, respectively, during decoding. An utterance-specific factor was estimated to 1.24.

Conclusion

A novel method to incorporate time-varying speaker characteristic properties into the acoustic model was presented and applied to vocal tract length normalization. The main performance improvement compared with the conventional utterance-specific warping is due to the factor decomposition into a static and a dynamic component. In this way, the dynamic range during an utterance is limited to what can be expected for an individual speaker. Of the tested conditions, the best position for implementing the dynamic component was between phone sub-states, keeping the warp factor fixed within each state. The method lowered the standard utterance-specific warping error rate by 10% relative.

We are optimistic regarding its potential for further improvement. The proposed technique is flexible and was straightforwardly implemented in a conventional phone-based HMM system. The rate-of-change constraints of the warp factor are efficiently implemented in the transition matrix. This approach makes it possible to train model-specific dynamic properties of the warp factor using conventional training procedures. Further work will be directed to such training.

The proposed approach is not limited to frequency warping, but can be applied to other time-varying speaker characteristic and environmental properties, such as voice source quality, speech rate, non-stationary background noise, microphone distance, etc.

Acknowledgement

The work was funded by the Swedish Research Council.

References

Akhil T, Rath P, Umesh S and Sanand D R (2008). A Computationally Efficient Approach to Warp Factor Estimation in VTLN Using EM Algorithm and Sufficient Statistics. *Proceedings of Interspeech*, 1713-1716.

Blomberg M and Elenius D (2007). Vocal tract length compensation in the signal and model domains in child speech recognition. *Proceedings of Fonetik-2007*. TMH-QPSR, 50(1), KTH, Stockholm.

Blomberg M and Elenius, D (2008). Investigating Explicit Model Transformations for Speaker Normalization. *Proceedings of ISCA-ITRW Speech Analysis and Processing for Knowledge Discovery*, Aalborg, Denmark.

Blomberg M and Elenius, D (2009). Tree-based estimation of speaker characteristics for speech recognition. *Proceedings of Interspeech*, 580-583.

Dusan S (2007). Vocal Tract Length during Speech Production. *Proceedings of Interspeech*, 1366-1369.

Elenius D (2010). *Accounting for Individual Speaker Properties in Automatic Speech Recognition*. Lic. Thesis, CSC/TMH, KTH, Sweden.

Elenius D and Blomberg M (2009). On Extending VTLN to Phoneme-specific Warping in Automatic Speech Recognition. *Proceedings of Fonetik 2009*. Dept. of Linguistics, Stockholm University.

Fant G (1960). *Acoustic Theory of Speech Production*. The Hague, The Netherlands: Mouton.

Fant G (1975). Non-uniform vowel normalization. *Quarterly Progress and Status Report*, Department for Speech Music and Hearing, Stockholm, Sweden.

Gales M J F (1998). Predictive model-based compensation schemes for rubust speech recognition. *Speech Communication*, 25: 49-74.

Giuliani D, Gerosa M and Brugnara F (2006). Improved Automatic Speech Recognition Through Speaker Normalization. *Computer Speech & Language*, 20(1): 107-123.

Lee L and Rose R C (1996). Speaker Normalization Using Efficient Frequency Warping Procedures. *Proceedings of ICASSP*, 353-356.

Maragakis G and Potamianos A (2008). Region-Based Vocal Tract Length Normalization for ASR. *Proceedings of Interspeech*, 1365-1368.

Miguel A, Lleida E, Rose R C, Buera L and Ortega A (2005). Augmented state space acoustic decoding for modeling local variability in speech. Proceedings of ICSLP, 3009-3012.

Pitz M and Ney H (2005). Vocal Tract Normalization Equals Linear Transform in Cepstral Space. *IEEE Transactions on Speech and Audio Processing*, 13(5): 930-944.

Potamianos A and Narayanan S (2003). Robust Recognition of Children's Speech. *IEEE Transactions on Speech and Audio Processing*, 11(6): 603-616.