On extending VTLN to phoneme-specific warping in automatic speech recognition

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

Abstract

Phoneme- and formant-specific warping has been shown to decrease formant and cepstral mismatch. These findings have not yet been fully implemented in speech recognition. This paper discusses a few reasons how this can be. A small experimental study is also included where phoneme-independent warping is extended towards phoneme-specific warping. The results of this investigation did not show a significant decrease in error rate during recognition. This is also in line with earlier experiments of methods discussed in the paper.

Introduction

In ASR, mismatch between training and test conditions degrades the performance. Therefore much effort has been invested into reducing this mismatch using normalization of the input speech and adaptation of the acoustic models towards the current test condition.

Phoneme-specific frequency scaling of a speech spectrum between speaker groups has been shown to reduce formant- (Fant, 1975) and cepstral- distance (Potamianos and Narayanan, 2003). Frequency scaling has also been performed as a part of vocal tract length normalization (VTLN) to reduce spectral mismatch caused by speakers having different vocal tract lengths (Lee and Rose 1996). However, in contrast to findings above this scaling is normally made without regard to sound-class. How come that phoneme-specific frequency scaling in VTLN has not yet been fully implemented in ASR (automatic speech recognition) systems?

Formant frequency mismatch was reduced by about one-half when formant- and vowel-category- specific warping was applied compared to uniform scaling (Fant, 1975). Also phoneme-specific warping without formant-specific scaling has been beneficial in terms of reducing cepstral distance (Potamianos and Narayanan, 2003). In the study it was also found that warp factors differed more between phonemes for younger children than for older ones. They did not implement automatic selection of warp factors to be used during recognition. One reason presented was that the gain in practice could be limited by the need of correctly estimating a large number of warp factors. Phone clustering was suggested as a method to limit the number of warping factors needed to estimate.

One method used in ASR is VTLN, which performs frequency warping during analysis of an utterance to reduce spectral mismatch caused by speakers having different vocal tract lengths (Lee and Rose, 1996). They steered the degree of warping by a time-independent warping-factor which optimized the likelihood of the utterance given an acoustic model using the maximum likelihood criterion. The method has also been frequently used in recognition experiments both with adults and children (Welling, Kanthak and Ney, 1999; Narayanan and Potamianos, 2002; Elenius and Blomberg, 2005; Giuliani, Gerosa and Brugnara 2006). A limitation with this approach is that time-invariant warping results in all phonemes as well as non-speech segments sharing a common warping factor.

In recent years increased interest has been directed towards time-varying VTLN (Miguel et.al., 2005; Maragakis et.al., 2008). The former method estimates a frame-specific warping factor during a memory-less Viterbi decoding process, while the latter method uses a two-pass strategy where warping factors are estimated based on an initial grouping of speech frames. The former method focuses on revising the hypothesis of what was said during warp estimation while the latter focuses on sharing the same warp factor within each given group. Phoneme-specific warping can be implemented to some degree with either of these methods. Either by explicitly forming phoneme-specific groups or implicitly by estimating frame-specific warp factors.

However, none of the methods above presents a complete solution for phoneme-specific warping. One reason is that more then one instantiation of a phoneme can occur far apart in time. This introduces a long distance depend-

ency due to a shared warping factor. For the frame-based method using a memory-less Viterbi process this is not naturally accounted for.

A second reason is that in an unsupervised two-pass strategy initial mismatch causes recognition errors which limit the performance. Ultimately initial errors in assigning frames to group-identities will bias the final recognition phase towards the erroneous identities assigned in the first pass.

The objective of this paper is to assess the impact of phoneme-specific warping on an ASR-system. First a discussion is held regarding issues with phoneme-specific warping. Then an experiment is set up to measure the accuracy of a system performing phoneme-specific VTLN. The results are then presented on a connected-digit task where the recognizer was trained for adults and evaluated on children's speech.

Phoneme-specific VTLN

This section describes some of the challenges in phoneme-specific vocal tract length normalization.

Selection of frequency warping function

In (Fant, 1975) a case was made for vowel-category and formant- specific scaling in contrast to uniform scaling. This requires formant tracking and subsequent calculations of formant-specific scaling factors, which is possible during manual analysis. Following an identical approach under unsupervised ASR would include automatic formant tracking, which is a non-trivial problem without a final solution (Vargas at.al, 2008).

Lee and Rose (1996) avoided explicit warping of formants by performing a common frequency warping function for all formants. Since the function is equal for all formants, no formant-frequency estimation is needed when applying this method. The warping function can be linear, piece-wise linear or non-linear. Uniform frequency scaling of the frequency interval of formants is possible using a linear or piece-wise linear function. This could also be extended to a rough formant scaling, using a non-linear function, under the simplified assumption that the formant regions do not overlap.

This paper is focused on uniform scaling of all formants. For this aim a piece-wise linear warping function is used, where the amount of warping is steered by a warping-factor.

Warp factor estimation

Given a specific form of the frequency warping to be performed, a question still remains of the degree of warping. In Lee and Rose (1996) this was steered by a common warping factor for all sound-classes. The amount of warping was determined by selecting the warping-factor that maximized the likelihood of the warped utterance given an acoustic model. In the general case this maximization lacks a simple closed form and therefore the search involves an exhaustive search on a set of warping factors.

An alternative to warp the utterance is to perform a transform of the model parameters of the acoustic model towards the utterance. Thereby a warp-specific model is generated. In this case, warp factor selection amounts to selecting the model that best fit data, which is a standard classification problem. So given a set of warp-specific models one can select the model that results in the maximum likelihood of the utterance.

Phoneme-specific warp estimation

Let us consider extending the method above to a phoneme-specific case. Instead of a scalar warping factor a vector of warping factors can be estimated with one factor per phoneme. The task is now to find the parameter vector that maximizes the likelihood of the utterance given the warped models. In theory this results in an exhaustive search of all combinations of warping factors. For 20 phonemes with 10 warp candidates, this amounts to 10^{20} likelihood calculations. This is not practically feasible and thereby an approximate method is needed.

In (Miguel et.al, 2005) a two-pass strategy was used. During the first pass a preliminary segmentation is made. This is then held constant during warp estimation to allow separate warp-estimates to be made for each phoneme. Both a regular recognition phase as well as K-Means grouping has been used in their region-based extension to VTLN

The group-based warping method above relies on a two-pass strategy where a preliminary fixed classification is used during warp factor estimation, which is then applied in a final recognition phase. Initial recognition errors can ultimately cause a warp to be selected that maximizes the likelihood of an erroneous identity. Application of this warping factor will then

bias the final recognition towards the erroneous identities. The severity of this hazard depends on the number of categories used and the kind of confusions made.

An alternative to a two-pass approach is to successively revise the hypothesis of what has been said as different warping factors are evaluated. Following this line of thought leads to a parallel of warp factor estimation and determination of what was said. For a speech recognizer using Viterbi-decoding this can be implemented by adding a warp-dimension to the phoneme-time trellis (Miguel et.al. 2005). This leads to a frame-specific warping factor. Unconstrained this would lead to a large amount of computations. Therefore a constraint on the time-derivative of the warp factor was used to limit the search space.

A slowly varying warping factor might not be realistic even though individual articulators move slowly. One reason is that given multiple sources of sound a switch between them can cause an abrupt change in the warping factor. This switch can for instance be between speakers, to/from non-speech frames, or a change in place and manner of articulation. The change could be performed with a small movement which causes a substantial change in the air-flow path. To some extent this could perhaps be taken into account using parallel warp-candidates in the beam-search used during recognition.

In this paper model-based warping is performed. For each warp setting the likelihood of the utterance given the set of warped models is calculated using the Viterbi algorithm. The warp set is chosen that results in the maximum likelihood of the utterance given the warped models.

In contrast to the frame-based, long distance dependencies are taken into account. This is handled by warping the phoneme models used to recognize what was said. Thereby each instantiation of the model during recognition is forced to share the same warping factor. This was not the case in the frame-based method which used a memory-less Viterbi decoding scheme for warp factor selection.

Separate recognitions for each combination of warping factors were used to avoid relying on an initial recognition phase, as was done in the region-based method.

To cope with the huge search space two approaches were taken in the current study namely: reducing the number of individual warp factors by clustering phonemes together and by supervised adaptation to a target group.

Experimental study

Phoneme-specific warping has been explored in terms of WER (word error rate) in an experimental study. This investigation was made on a connected-digit string task. For this aim a recognition system was trained on adult speakers: This system was then adapted towards children by performing VTLT (vocal tract length transformation). A comparison between phoneme-independent and -specific adaptation through warping the models of the recognizer was conducted. Unsupervised warping during test was also conducted using two groups of phonemes with separate warping factors. The groups used were formed by separating silence, /t/ and /k/ forming the rest of the phonemes.

Speech material

The corpora used for training and evaluation contain prompted digit-strings recorded one at a time. Recordings were made using directional microphones close to the mouth. The experiments were performed for Swedish using two different corpora, namely SpeeCon and PF-STAR for adults and children respectively.

PF-STAR consists of children speech in multiple languages (Batliner et.al. 2005). The Swedish part consists of 198 children of 4 to 8 years repeating oral prompts spoken by an adult speaker. In this study only connected-digit strings were used to concentrate on acoustic modeling rather than language models. Each child was orally prompted to speak 10 three-digit strings amounting to 30 digits per speaker. Recordings were performed in a separate room at daycare and after-school centers. During these recordings sound was picked up by a head-set mounted cardioid microphone, Sennheiser ME 104. The signal was digitized using 24 bits @ 32 kHz using an external usb-based A/D converter. In the current study the recordings were down-sampled to 16 bits @ 16 kHz to match that used in SpeeCon.

SpeeCon consists of both adults and children down to 8 years (Großkopf et.al 2002). In this study, only digit-strings recordings were used. The subjects were prompted using text on a computer screen in an office environment. Recordings were made using the same kind of microphone as was used in Pf-Star. An analog high-pass filter with a cut-off frequency of 80 Hz was used, and digital conversion was performed

using 16 bits at 16 kHz. Two sets were formed for training and evaluation respectively consisting of 60 speakers each to match Pf-Star.

Recognition system

The adaptation scheme was performed using a phone-level HMM-system (Hidden Markow Model) for connected digit-string recognition. Each string was assumed to be framed by silence (/sil/) and consist of an arbitrary number of digit-words. These were modeled as concatenations of three state three-phone models ended by an optional short-pause model. The short pause model consisted of one state, which shared it's pdf (probability density function) with the centre state of the silence model.

The distribution of speech features in each state was modeled using GMMs (Gaussian Mixture Models) with 16 mixtures and diagonal covariance matrices. The feature vector used consisted of 13 * 3 elements. These elements correspond to static parameters and their first and second order time derivatives. The static coefficients consisted of the normalized log energy of the signal and MFCCs (Mel Frequency Cepstrum Coefficients). These coefficients were extracted using a cosine transform of a mel scaled filter bank consisting of 38 channels in the range corresponding to the interval 0 to 7.6 kHz.

Training and recognition experiments were conducted using the HTK speech recognition software package (Young et.al., 2005). Phoneme-specific adaptation of the acoustic models and warp factor search was performed by separate programs. The adaptation part was performed by applying the corresponding piece-wise linear VTLT in the model space as was used in the feature space by Pitz and Ney 2005.

Results

The WER (word error rate) of recognition experiments where unsupervised adaptation to the test utterance was performed is shown in Table 1. The baseline experiment using phoneme-independent warping resulted in a WER (word error rate) of 13.2%. Introducing two groups ({/sil/, /t/, /k/} and {the rest of the models}) with separate warping factors lowered the error rate to 12.9%. This required that an exhaustive search of all combinations of two warping factors was performed. If an assumption that the warping factor could be estimated

separately, the performance increase was reduced by 0.2% absolute. Further division by forming a 3:rd group with unvoiced fricatives {/s/, /S/, /f/ and /v/} was also attempted, but with no improvement in recognition to that above. In this case /v/ in "två" is mainly unvoiced

Table 1. Recognition results with model group-specific warping factors. Unsupervised likelihood maximization of each test utterance. The group was formed by separating /sil/, /t/ and /k/ from the rest of the models.

Method	WER
VTLN 1-warping factor	13,2
Speech	13,4
2 Groups (separate estimation)	13,1
2 Groups (joint maximization)	12,9

Phoneme-specific adaptation of an adult recognizer to children resulted in warping factors given in Figure 1. The method gave silence a warping factor of 1.0, which is reasonable. In general voiced-phonemes were more strongly warped than un-voiced ditto.
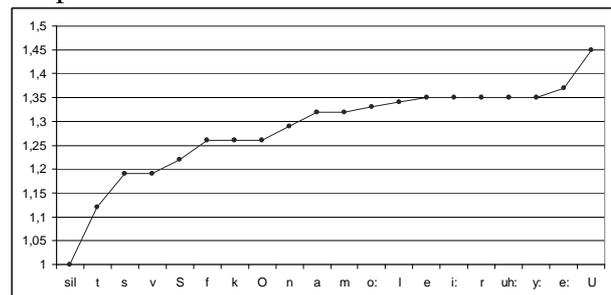

Figure 1. Phoneme-specific warp adapting adult models to children sorted in increasing warp-factor.

Further division of the adaptation data into age groups resulted in the age and phoneme-specific warping factors shown in Figure 2. In general, the least warping of adult models was needed for 8 year old children compared to younger children.
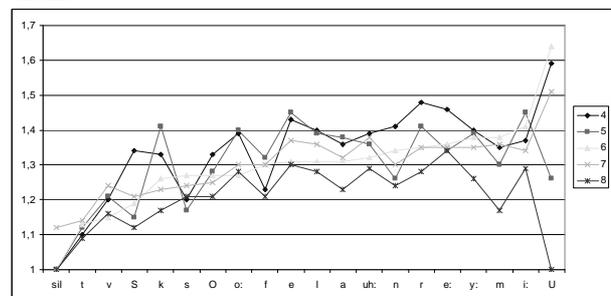

Figure 2. Phoneme and age-specific warping factors. Optimized on likelihood of adaptation data. The phonemes are sorted in increasing warp-factor for 6 year old speakers.

The found warping factors for the child and age groups were then applied on the test-data to measure the implication on the WER. The result of this experiment is given in Table 2. Introducing phoneme-specific warping did not substantially reduce the number of errors compared to a shared warping factor for all phonemes.

Table 2. Recognition results with adult model adapted to children using a fixed warping vector for all utterances with one warp factor per phoneme .Phoneme- dependent and -independent warping is denoted Pd and Pi respectively.

Method	WER
Fix Pi	13,7
Fix Pd	13,2
Fix Pd per age	13,2

Discussion

Time invariant VTLN has in recent years been extended towards phoneme-specific warping. The increase in recognition accuracy during experimental studies has however not yet reflected the large reduction in mismatch shown by Fant (1975).

One reason for the discrepancy can be that unconstrained warping of different phonemes can cause unrealistic transformation of the phoneme space. For instance swapping places of the low left and upper right regions could be performed by choosing a high and low warping factor respectively.

Conclusion

In theory phoneme-specific warping has a large potential for improving the ASR accuracy. This potential has not yet been turned into significantly increased accuracy in speech recognition experiments. One difficulty to manage is the large search space resulting from estimating a large number of parameters. Further research is still needed to explore remaining approaches of incorporating phoneme-dependent warping into ASR.

Acknowledgements

References

Batliner A, Blomberg M, D'Acry S, Elenius D and Giuliani D. (2005). The PF_STAR Children's Speech Corpus. Interspeech 2005, 2761 – 2764.

Elenius, D., Blomberg, M. (2005) Adaptation and Normalization Experiments in Speech Recognition for 4 to 8 Year old Children. In Proc Interspeech 2005, pp. 2749 - 2752.

Fant, G. (1975) Non-uniform vowel normalization. STL-QPSR. Quartely Progress and Status Report. Departement for Speech Music and Hearing, Stockholm, Sweden 1975.

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

Großkopf B, Marasek K, v. d. Heuvel, H., Diehl F, and Kiessling A (2002). SpeeCon - speech data for consumer devices: Database specification and validation. Second International Conference on Language Resources and Evaluation 2002.

Lee, L., and Rose, R. (1996) Speaker Normalization Using Efficient Frequency Warping Procedures. In proc. Int. Conf. on Acoustic, Speech and Signal Processing,1996, Vol 1, pp. 353-356.

Maragakis, M. G. and Potamianos, A. (2008) Region-Based Vocal Tract Length Normalization for ASR. Interspeech 2008. pp. 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. In Proc. Int. Conf. Spoken Language Processing. Sep 2005.

Narayanan, S., Potamianos, A. (2002) Creating Conversational Interfaces for Children. IEEE Transactions on Speech and Audio Processing, Vol. 10, No. 2, February 2002.

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

Potamianos, A. Narayanan, S. (2003) Robust Recognition of Children's Speech. IEEE Transactions on Speech and Audio Processing, Vol 11, No 6, November 2003. pp. 603 – 616.

Welling, L., Kanthak, S. and Ney, H. (1999) Improved Methods for Vocal Tract Normalization. ICASSP 99, Vol 2, pp. 161-164.

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P. (2005) The HTK book. Cambridge University Engineering Department 2005.

Vargas, J. and McLaughlin, S. (2008). Cascade Prediction Filters With Adaptive Zeros to Track the Time-Varying Resonances of the Vocal Tract. In Transactions on Audio Speech, and Language Processing. Vol. 16 No 1 2008. pp. 1-7.