Model space size scaling for speaker adaptation

Mats Blomberg
Speech, Music and Hearing, KTH/CSC, Stockholm

Abstract

In the current work, instantaneous adaptation in speech recognition is performed by estimating speaker properties, which modify the original trained acoustic models. We introduce a new property, the size of the model space, which is included to the previously used features, VTLN and spectral slope. These are jointly estimated for each test utterance. The new feature has shown to be effective for recognition of children's speech using adult-trained models in TIDIGITS. Adding the feature lowered the error rate by around 10% relative. The overall combination of VTLN, spectral slope and model space scaling represents a substantial 31% relative reduction compared with single VTLN. There was no improvement among adult speakers in TIDIGITS and in TIMIT. Improvement for this speaker category is expected when the training and test sets are recorded in different conditions, such as read and spontaneous speech.

Introduction

In this paper, initial work is presented on including a new speaker property for speaker adaptation. This property is the size of the space spanned by the set of trained models. It is well known that this type of acoustic property is closely related to articulation clarity, speech rate, and to difference between speech styles, such as read and spontaneous speech (e.g. Lindblom, 1963, Nord, 1986). There are also indications that the reduced spectral space of spontaneous speech in comparison with read speech is a major cause of the decrease of recognition accuracy in spontaneous speech (Nakamura, Iwano and Furui, 2008). These findings support the hypothesis that adaptation to this property would improve recognition performance in these types of mismatch.

In previous work, we have used vocal tract length and spectral slope for instantaneous speaker adaptation (Blomberg & Elenius, 2008, 2009). In the current paper, model space scaling is jointly estimated with these properties

Model space scaling

We apply the procedure, in a framework where the speaker properties are estimated by maximizing the likelihood output of the recognizer on the test utterance. In this search, the property values are implemented by property-specific transformations on the trained models and a recognition procedure is performed for each examined value. An alternative to transforming the models would be to perform the inverse transformation on the test utterance. We have chosen to operate on the models, since this facilitates phoneme-specific transformation.

The transformation implements a simple radial movement of the mean vector of each mixture component in a set of continuous-density HMMs towards/away-from a center-of-gravity point which is common to all models in the set. The new position of a component is derived by scaling its distance to the center-of-gravity by a scaling factor.

For a model space scaling factor α, $0 < \alpha$, the scaled mean feature vector of a mixture component will be

$$\begin{aligned}\tilde{\mu}_{ijk} &= CG_H + \alpha(\mu_{ijk} - CG_H) = \\ &= \alpha u_{ijk} + (1-\alpha) CG_H \ , \end{aligned} \qquad (1)$$

where u_{ijk} is the average feature vector of mixture component number k in state number j of model number i and CG_H is the center-of-gravity of the model set H. A scale factor value $0 < \alpha < 1$ corresponds to a compression of the model space. $\alpha > 1$ corresponds to an expansion. This linear equation is basically the same as one which was used to map formant frequencies of short vowels in mono-syllabic words to those spoken in sentences (Stålhammar, Karlsson and Fant, 1973).

It should be noted that linear scaling in the spectral or cepstral domains will not give the same result as in the formant frequency domain. Furthermore, studies have mainly been performed on vowels and the function for consonants is not as well known. For these reasons, it is uncertain if the simple linear interpolation formula in Eq. (1) will model the actual relations accurately enough to improve recognition performance.

It is possible to scale the static and the time differential elements differently. Even if both categories may be important for scaling, it is quite likely that the best scale factor value differs between them. It may therefore be necessary to estimate separate values for these feature categories.

The speech rate affects both static and dynamic features and is consequently expected to have impact on the model space. For this reason, it may be of interest to use speech corpora with mismatch in this respect for experiments. Children's speech has been found to be generally slower compared with adults (Lee, Potamianos and Narayanan, 1999) and is therefore a good candidate for evaluation. It would also be interesting to study read vs. spontaneous speech. This is planned for future work.

Three variance scaling functions have been considered. These are: (i) no change, (ii) the same scale factor as for the mean values and (iii) squared mean scale factor. In preliminary experiments, the best performance was achieved when the variance was not changed. This was used for the subsequent experiments in the paper. Further studies are required for a more decisive conclusion.

Experiments

In the experiments performed, model space size is evaluated in combination with frequency warping (Vocal Tract Length Normalization, VTLN) (Lee and Rose, 1996) and spectral slope. A low number of values of each property are examined in all combinations with the other properties. In these preliminary experiments, the model space scaling factors were tentatively set to 8 values from 0.8 through 1.5 with a linear step of 0.1. The frequency warping factor was quantized into 16 log-spaced values between 0.8 and 1.7 in TIDIGITS and between 0.79 and 1.24 in TIMIT. Spectral slope was implemented by two parameters, a spectral real pole and a spectral real zero. The pole and zero cut-off frequencies were varied in 8 logarithmically spaced steps between 100 and 4000 Hz (Blomberg and Elenius, 2009).

Corpora

Two American-English corpora, TIDIGITS and TIMIT, were chosen for initial evaluation. TIDIGITS consists of digit strings spoken by adults and children of both genders. The adult test set consists of 28583 digits. The adult male, the adult female and the children's test sets contain 14159, 14424, and 12637 digits, respectively. Models were trained on two sets: the adult (male + female) and the adult male training speakers. Evaluation was performed for the separate adult, male, female and children's test sets.

TIMIT contains read sentences of 630 adult speakers. The training set consists of 4620 utterances spoken by 462 subjects. The full test set of 1344 sentences from 168 speakers was used for evaluation.

System

The TIDIGITS experiments were performed using a connected-digit recognition system with triphone HMMs implemented in HTK. In TIMIT, monophones and a phoneme pair grammar (equal probabilities) were used.

In both cases, the acoustic models had 3 states with GMMs consisting of 32 mixture components and diagonal covariance matrices. Models were trained with a 57-dimensional acoustic feature vector, composed by 18 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 and truncation from 18 cepstral coefficients to 12 after transformation as in (Blomberg and Elenius, 2008). A standard 39-element feature vector was, thus, used in the decoder.

To reduce the computational load of searching the very large space of speaker property values, the estimation was performed by a tree-based joint search algorithm

(Blomberg & Elenius, 2009). In this procedure, an iterative recognition search starts at the root of the tree, which contains broad models representing all allowed values of the speaker properties. Child node models each represent a subset of the mother node property values. The maximum scoring child node for the test utterance is selected for further search until a leaf node is reached, whose corresponding models represent a single value of each property.

In the absence of separate development data, the insertion likelihood ("penalty") was adjusted to minimize the error rate on the baseline case of adult test data using the original adult model.

Results and Discussion

Results on TIDIGITS for varying sets of speaker properties and combinations of training and test speaker categories are presented in Table 1. Adding model space size adaptation to VTLN and spectral slope reduces the error rate by around 10% relative for children using adult or male models. The improvement when including this property indicates that there is a systematic difference between adult and child speech in the size of their spectral space and that the proposed technique can compensate for this. The spectral space difference agrees with (Lee, Potamianos and Narayanan, 1999).

There is no such error reduction visible between any of the adult speaker categories. A possible interpretation is that there is no space size mismatch between the two categories male and female speakers. Even though there may be differences between individual adult speakers in this respect, this variability is already included in the training data.

In order to have an indication of which elements of the acoustic feature vector that are mainly involved in the improvement with model space size adaptation, we ran two new experiments for children's speech against male adult models. The experiments differed from the previous ones in that only the static or the time differential elements were adapted. The results are presented in Table 2.

When model space size was performed only on the static elements of the feature vector, the error rate was not reduced compared with no size adaptation. When instead adapting only the time differential elements, the error rate decreased compared with adapting both static and dynamic elements. These results show clearly that it was the dynamic properties, which reduced the error rate by this kind of adaptation.

Even lower error rate was achieved by another selection criterion in the hierarchical search tree. When the model with the highest likelihood along the search path was chosen, the error rate was lowered further to 2.09%.

The distribution of the estimated size scale factor in the adult-male/child case and when only the time differential feature elements are adapted is displayed in Figure 1. For a majority of the utterances, the model space is compressed. Evidently, children's speech has in general slower and smoother transitions than that of adult males. This is in agreement with previous findings that children's speech is slower than that of adults (Lee, Potamianos and Narayanan, 1999). It is also obvious that the minimum allowed factor value has been set too high. Still better results are expected when this will be corrected in further experiments.

Table 1. WER for adaptation to different sets of speaker properties in TIDIGITS. Model space scaling is denoted "Size".

Train set	Adult	Adult	Adult	Adult	Male	Male	Male	Male
Test set	Adult	Male	Female	Child	Adult	Male	Female	Child
Original	0.55	0.76	0.35	3.17	6.44	0.58	12.19	46.73
Size	0.55	0.76	0.35	2.99	5.84	0.56	11.02	44.76
Slope	0.53	0.73	0.33	2.95	5.51	0.58	10.41	42.41
VTLN	0.54	0.73	0.35	1.23	0.64	0.52	0.76	3.56
Slope+Size	0.52	0.73	0.33	2.65	5.00	0.57	9.35	40.11
VTLN+Size	0.54	0.72	0.35	1.20	0.64	0.54	0.74	3.36
VTLN+Slope	0.55	0.74	0.36	1.07	0.62	0.52	0.66	2.83
VTLN+Slope+Size	0.54	0.74	0.35	0.96	0.63	0.54	0.71	2.58

Table 2. Word error rate with size estimation of different parts of the acoustic feature vector. Training and test speakers were male adults and children, respectively.

No size adaptation	2.83
All features adapted	2.58
Only static features adapted	2.85
Only Delta+Accel. features adapted	2.46

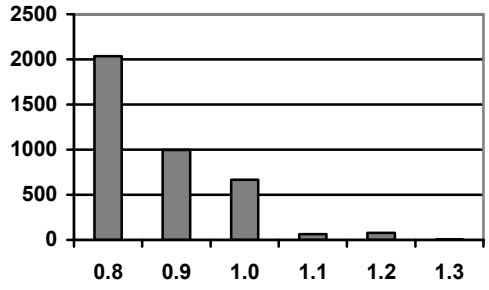

Figure 1. Histogram of number of utterances with estimated model space scale factor for time-differential acoustic features in children's speech using male adult models.

A few experiments have been performed on TIMIT. Due to computational load considerations, spectral slope was excluded from adaptation. The results are shown in Table 3.

Table 3. TIMIT results (Phoneme Error Rate). Both static and dynamic features are adapted.

Baseline	VTLN	VTLN+Size
37.03	36.66	36.64

In the adult/adult condition of TIMIT, there is a small improvement from VTLN but no further improvement from model space scaling, similarly to TIDIGITS. We tried scaling only static or differential features as well as estimating different scale factors for the two feature groups. These settings had only marginal influence on the result. A likely explanation to the lack of improvement in TIMIT is that there is no speaker mismatch between training and test speakers in the model space size respect.

Conclusions

Speaker adaptation by adjustment of model set size is efficient for the recognition of children's speech using adult or male adult models. Adding model space size to vocal tract length and spectral slope in a joint estimation framework lowered the word error rate by 10% and 9% relative, respectively, for the two training speaker categories. When also excluding static features from adaptation, the error rate using male adult models was further decreased by 5% relative. This overall combination of VTLN, spectral slope and model space scaling represents a substantial 31% relative reduction compared with single VTLN.

The method needs to be further developed for better scaling of the static features. For vowels, scaling in the formant frequency domain would be a natural choice, but the theoretical advantage is reduced by the unavoidable formant tracking errors.

Still another possibility would be to allow time varying model space size, as has been done for VTLN (Elenius and Blomberg, 2010).

Further experiments include testing on corpora with speech style mismatch between training and test, such as between read and spontaneous speech.

References

Blomberg, M, and Elenius, D (2008). Investigating explicit model transformations for speaker normalization. *Proc. of ISCA ITRW Speech Analysis and Processing for Knowledge Discovery*.

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

Elenius D and Blomberg, M (2010). Dynamic vocal tract length normalization in speech recognition. *Proc. of Fonetik 2010*. Centre for Languages and Literature, Lund University. 29-34.

Lee S, Potamianos A, and Narayanan S (1999). Acoustics of children's speech: Developmental changes of temporal and spectral parameters. *J. Acoust. Soc. Amer.* 105: 1455–1468.

Lee L and Rose R C (1996). Speaker normalization using efficient frequency warping procedures. *Proc. of ICASSP*. 353-356.

Lindblom, B (1963) Spectrographic study of vowel reduction. *J. Acoust. Soc. Am.* 35:1773-1781.

Nakamura M, Iwano K, and Furui S (2005). Analysis of Spectral space reduction in spontaneous speech and its effects on speech recognition performances. *Proc. of Interspeech 2005*, 3381-3384.

Nord L (1986). Acoustic studies of vowel reduction in Swedish. *STL-QPSR* 27/4: 19-36.

Stålhammar U, Karlsson I, Fant G (1973). Contextual effects on vowel nuclei. *STL-QPSR* 14/4: 1-18.