

MULTIPLE ACOUSTIC AND VARIABILITY ESTIMATION MODELS FOR ASR

Stéphane Dupont and Christophe Ris

Multitel, Avenue Copernic 1, B-7000 Mons, Belgium
{dupont,ris}@multitel.be

ABSTRACT

In the paper, we expose a formalism that allows to make use of features representing both short-term and long-term speech behavior. This amounts to using multiple (specific, compensated or adapted) acoustic models which are defined according to additional hidden variables not pertaining to the phonetic sequence, but rather to long-term stable structures in the speech signal, like the speaker identity or the speaking rate.

This formalism has been evaluated for recognition using vocal tract length (VTL) normalization. Features based on long-term pitch and formant measures, as well as PCA reductions of these, have been investigated and show significant correlation with the VTL. Speech recognition experiments performed on the children portion of the TI-DIGITS database show the improved accuracy obtained using this technique compared to VTL selection based on the traditional Maximum Likelihood criterion.

1. INTRODUCTION

Vocal tract length normalization [1] is a popular technique that allows to obtain features that are less dependent on the inter-speaker spectral differences. In this framework, different techniques have been shown appropriate for estimating an optimal warping factor. Generally, the procedure relies on maximum likelihood estimation. Feature vector sequences obtained using a range of different spectral warping factors are used to generate different recognition hypotheses. The warping factor that provides the hypothesis that has maximum likelihood is then selected. This procedure involves running several decoding processes in parallel, which in a straightforward implementation, multiplies the computations by the number of warping factors that are explored. Pruning the search space across those different decoding may hence be beneficial. In [2], the decoding further permits frame-specific warping functions by allowing transitions from states related to different warping factors. Constraints (reflecting physiologically plausible degrees of variability) can be applied in this process, for instance so that the frequency warping transformations applied to adjacent frames are using adjacent warping factors. The baseline algorithm described before is a particular case where transition from one warping to

another is not allowed within a speech utterance to be recognized.

A second strategy, based on maximum-likelihood consists in first running a decoding to obtain an optimal state-path with no warping applied. the likelihood for different warping factors applied is then computed and the optimal warping selected. Features computed based on this warping estimate are then used in a second-pass decoding. This approach may also be suboptimal, especially in situations where the first-pass model is far from matching the test data (in the case of children speech for instance).

These VTLN approaches finally consist in designing multiple feature extraction/acoustic models couples and selecting the most appropriate at test time, using a likelihood criterion. Other criteria have also been used with some success to define multiple models that are then competing during test time, for instance gender-dependent models [3] and speaking rate dependent models [4].

A criticism that may apply to these maximum likelihood approaches is that they purely rely on the feature sets and acoustic models that have been designed for phonetic (or allophonic) classification. These modeling structures, and especially the feature sets that are used as well as the underlying independence assumptions that are applied, are the result of years of research in optimizing the phonetic and word classification accuracies which, understandably, mainly rely on local representations of the signal. Estimating factors such as the VTL, the gender, the speaking rate, or other variability of the speech signal may however benefit from longer term and/or alternate acoustic representations, such as average pitch [5] and formant frequencies in the case of VTL, or longer-term modulation spectra [6] in the case of speaking rate. The choice of speech analysis is known to have a strong impact on phoneme and word recognition accuracy. The same is likely to be true for the estimation of factors reflecting secondary (non-linguistic) sources of speech variability, like the vocal tract length or the speaking rate. In the following section, we present a formalism that introduces those variabilities as additional hidden variables, for which the model structure and the feature extraction approach is further subject to research.

Later on, this structure is applied to speech recognition using a partitioning in terms of VTL, and experimental results

are presented on the TI-DIGITS corpus.

1.1. Modeling of Hidden Speech Variability

The task of speech recognition based on the acoustic signal O consists in estimating the most likely hidden word sequence \widehat{M} according to:

$$\widehat{M} = \arg \max_M P(M|O) \quad (1)$$

$$= \arg \max_M P(O|M)P(M) \quad (2)$$

The first term is the one we are considering for acoustic modeling. This term is computed as the sum of the contributions of different state paths Q , or as the contribution of the best state-path Q (Viterbi assumption), as follows:

$$P(O|M) = \max_Q P(O, Q|M) \quad (3)$$

Besides introducing this hidden state sequence, we may also introduce an additional hidden variable C , representing longer-term phenomena such as the vocal tract length, the gender, the speaking style (or mode) or other:

$$\begin{aligned} P(O|M) &= \max_{Q, C} P(O, Q, C|M) \\ &= \max_{Q, C} P(Q|M)P(O, C|Q, M) \end{aligned} \quad (4)$$

The first term is the state sequence probability (estimated from the pronunciation models and HMM transition probabilities). We may then introduce several feature streams that represent the acoustics O , as well as several assumptions related to those. On one side we have X , the regular (frame-based, phonetically-relevant) features. On the other side, we introduce Y , the long-term features. Being designed to represent long-term properties of the signal, Y are assumed to be conditionally independent on the state sequence Q . Also C is assumed to be conditionally independent on the state sequence Q . From eq. 4, this leads us to:

$$\begin{aligned} P(O, C|Q, M) &= P(X, Y, C|Q, M) \\ &= P(C|Q, M)P(X, Y|Q, M, C) \\ &= P(C|Q, M)P(X|Q, C, M, Y)P(Y|Q, M, C) \\ &= P(C|M)P(X|Q, C, M, Y)P(Y|M, C) \\ &= P(Y|M)P(X|Q, C, M, Y)P(C|M, Y) \end{aligned} \quad (5)$$

The first term is the prior probability of the long term features and can be ignored as it is constant for a given utterance. Besides this term, this factorization of the objective function now brings two different acoustic terms on which different contradictory assumptions may be relevant. As in [7], we can further assume that the dependence of X on Y is fully accounted by the hidden state C . We hence have:

$$P(X|Q, C, M, Y) = P(X|Q, C, M) \quad (6)$$

Finally, in a more general way, eq. 5 may be written:

$$\begin{aligned} P(O, C|Q, M) &\simeq P(X|Q, C, M)^\beta P(C|M, Y)^{1-\beta} \end{aligned} \quad (7)$$

Where β can be tuned for optimal performance.

This formalism opens research opportunities in exploring features that are characteristic of longer term phenomena and of speaker characteristics. For instance, average pitch and formant measures over a speech segment or utterance convey important information about the speaker (its gender and age). These measures are very noisy given the HMM states used in speech recognition (basically phoneme variants), so that locally, such measures do not convey useful information for increasing phonetic discrimination.

1.1.1. Acoustic modeling term

For estimating $P(X|Q, C, M)$, the traditional (well understood) HMM assumptions and feature extraction methodologies can be used. The term corresponds to the usual acoustic likelihood, with an additional dependence on C . This can be estimated using multiple factorized acoustic models, one for each discretized value of C . Alternatively, if a model reflecting the effect of C on the acoustic features is known, it can be applied within a feature normalization scheme. This is for instance the case when C is considered to represent the speaker vocal tract length (VTL).

The assumption of independence of successive feature frames x_n (i.i.d assumption - independent and identically distributed) is one of the commonly accepted simplifications that are made in order to factorize the computation of this acoustic term:

$$P(X|Q, C, M) = \prod_n p(x_n|q_n, C, M) \quad (8)$$

Where n is the index of the acoustic frame, computed every 10 ms, and N the number of frames within the considered utterance. In this expression, local dependencies are usually accounted by using delta features or contextual frames (up to several hundred milliseconds [8]) while longer-term dependencies are discarded. These long-term dependencies are however mostly related to hidden modes, as indeed, speech frames are dependent on distant frames, for instance if these have been produced by the same speaker or expressed at similar speaking rate. The long-span dependencies of speech frames may hence be accounted by the C variable, and the i.i.d. assumption is hence expected to be less detrimental with appropriate C . This justifies the use of such modeling structures, where the models are conditioned on additional long-term hidden variables.

1.1.2. Variability estimation term

For the second term $P(C|M, Y)$, alternate modeling structure and feature extraction methodologies may be preferable. More particularly, as we consider that C represents secondary phenomena that are more stable than the non-stationary speech signal, the features representative of long-term signal portions

may be appropriate. We propose to define C using the following classes, which have been shown to significantly affect speech recognition accuracy [9]:

- Classes of speaking rate, for instance slow, medium and fast speech: in that case, appropriate features could be based on published ROS estimation methods [4], as well as modulation spectra [6].
- Classes expressing the spectral variability, such as the vocal tract length: in that case, pitch and formant measures may be appropriate. In this paper, we explore the use of long-term average measure of pitch and formants. The dimensionality of the feature vector obtained using those different measures is then reduced using PCA, which then allows to build a discrete model for estimating the probability of different VTL of a given utterance.
- Classes of foreign or regional accents: in that case, the assumption that leads to $P(Y|M, C)$ (from $P(Y|Q, M, C)$) looks more problematic. Indeed, the dependency of foreign accent realizations with respect to the state (phoneme) sequence is clearly significant, and hence, designing feature extraction schemes that allow to obtain auxiliary features Y that are independent on Q , but dependent on the accent, may be difficult. Note that this assumption also has to be validated in the previously considered cases. For instance, the pitch and formant frequencies should be made as independent as possible on the state sequence Q by integrating the local estimates on several utterances and by discarding the unvoiced portions from the estimation. Table 1 illustrates this point as, for instance, per-utterance estimation may not be stable enough to validate this assumption.

1.1.3. Discussion

Model structures based on a similar philosophy have been proposed in the literature. This approach is similar to the methods described in [10]. In that paper, auxiliary variables are introduced. These variables are presented as being either observed (they are hence considered as additional acoustic features for which appropriate modeling is sought) or else hidden (in which case the recognition simply relies on traditional MFCC/PLP features). Also, the additional (auxiliary) features are mostly computed at the local level. In our approach, additional hidden variables are introduced for which appropriate modeling and long-term feature analysis schemes are sought. This hence opens the research scope to defining these additional hidden variables, as well as devising feature extraction schemes that may be beneficial in modeling the phenomena represented by these additional variables.

Our model is more closely related to the mixture of expert (MofE) approach, where a 'gating network' estimates the reliability of each expert in an ensemble of models. The MofE approach has been reviewed and developed further in [7] where a state based formalism is proposed for introducing hidden variables and cooperative observations in the HMM equations. Unlike the approach proposed in this paper, the mixture of experts formalism applies at the HMM state level, so that the same assumptions are used for both HMMs and MofE-based models.

We believe that one of the key research issues related to this model, once the model structure and the nature of the C have been determined, consists in developing features Y that have a high correlation with the hidden value of C . In this paper, we consider the case of the vocal tract length. We exploit the long term average estimations of the formant and pitch frequencies to estimate $P(C|Y, M)$, in which C is quantized into 20 discrete warping values. The probability of other secondary phenomena, such as the speaking rate, or an accent, may require totally different acoustic representations. This model may hence allow to benefit from longer-term feature representations, which have rarely been shown to yield improvements in the traditional acoustic modeling framework.

2. VOCAL TRACT LENGTH

In this section, we are investigating the use of this approach in the case C is representing the warping factor (traditionally denoted α) adapted to the Vocal Tract Length of the speaker. The first acoustic term of the ASR likelihood function ($P(X|Q, C, M)$) corresponds to the usual Maximum Likelihood approach of VTLN. This is described in section 2.1. The second term ($P(C|Y, M)$) implies the estimation of different features and an additional model based on those features. This is described in section 2.2.

2.1. VTL normalization

The VTLN is implemented in the feature extraction module. Features used are PLP and the filter-bank is modified with a piece-wise linear warping method. In our experiments (see 3), we focus on the adaptation of children corpus to adults corpus. Therefore values of the warp scale factor C are inferior to 1.

The adaptation is also performed on the training set (adaptive training) for which "normalized" features are computed with the optimal C for each adult speaker. The resulting normalized acoustic model is more sensitive to the warping factor C but training data are more consistent, to the age variability point of view. Other specificities of our implementation are described in [11].

	F0	F1	F2	F3	Y
per-speaker	-0.442	-0.688	-0.851	-0.678	0.860
per-utterance	-0.248	-0.239	-0.250	-0.480	0.494

Table 1. Correlation between optimal VTL warping factors and long-term pitch and formant estimates. Y is the 1-dimension PCA reduced observation.

The local acoustic likelihoods $p(x_n|q_n, C, M)$ can hence be estimated for several discrete values of C using the VTL normalized features and the VTL-adapted acoustic model.

2.2. Features and modeling for VTL estimation

Long term average of the pitch ($F0$) and the first three formant ($F1, F2, F3$) estimates have been considered as long-term features for modeling the hidden VTL mode (i.e. $Y = [F0, F1, F2, F3]$). Although local pitch and formants values convey strong information on the prosody and the phonetic content of a sentence, averaging those values on several utterances shows a strong correlation with speaker specific indicators such as VTL warping factors (Table 1). Note that unvoiced portions of the speech signal are ignored.

Different models may be considered for the estimation of $P(C|M, Y)$. Artificial neural networks have powerful classification capabilities but require a lot of training observations. Alternatively, Gaussian mixture models may be used to estimate $P(Y|M, C)$. In the following experiments, facing the lack of training data, we have opted for a one-dimension discrete model where the unique long-term observation Y is obtained by PCA reduction of the four dimension vector $[F0, F1, F2, F3]$ to a single value.

Table 1 illustrates the correlation between the optimal warping factors (based on recognition accuracy) and the long-term pitch and formant estimates for children recordings. Besides the large difference between the per-speaker and per-utterance integrations, it is interesting to note that the best correlation occurs for $F3$ for the per-utterance case while it occurs for $F2$ for the per-speaker integration. It is known that the vowel class information is mostly conveyed by the pair $F1 - F2$ while $F3$ is more stable for a given speaker [12]. It is therefore likely that the per-utterance integration still depends on the phoneme sequence, somewhat invalidating the hypothesis discussed in section 1.1.2.

In the following experiments, the Y values have been discretized on 12 points Y_k , while 20 warping factors ranging from 0.64 to 1.0 are covering the variability of children voices. In practice, as in [5], for each Y_k intervals, a discrete pdf is built from the observation likelihoods of the warped utterances:

$$P(C|Y_k) = \frac{\sum_u I_u(C, Y)}{\sum_u \sum_j I_u(C_j, Y)}, \text{ U is the total nr. of utterances}$$

$$\text{Where } I_u(C, Y) = \begin{cases} P(C|X_u, M), & \text{if } Y = Y_k \\ 0, & \text{otherwise} \end{cases}$$

$$\text{And } P(C|X_u, M) = \frac{P(X_u, C|M)}{\sum_j P(X_u, C_j|M)}$$

$P(X_u, C|M)$ is the acoustic likelihood obtained by Viterbi alignment of the training utterances u for the each discrete value of C .

3. EXPERIMENTS

The database used is TI-Digits. This dialectically balanced database consists of more than 25,000 digit sequences spoken by 326 men, women, and children. The data were collected at 20 kHz in a quiet environment. For our experiments, all files are downsampled to 16 kHz.

The acoustic model is estimated using the adult training part of the corpus. The variability estimation model is estimated using the children training part of the corpus. Tests are performed on the children test part.

3.1. Modeling

The acoustic model is a multi-layer perceptron (MLP) trained on VTL normalized PLP features to classify into 22 context independent phonemes. The HMM have a left-to-right topology with a minimum state duration constraint computed as the half of the average duration of each phoneme. Note that the model is not specifically tuned to the connected digits recognition task (no word-models).

The long-term hidden variable C modeling is done according to section 2.2. A 12×20 discrete model has been built on the training portion of the TI-Digits children database. In order to have enough examples, averaging of pitch and formants as well as optimal warping estimation are performed utterance by utterance (yielding 3926 (C, Y) pairs).

3.2. Results

The children test set of TI-Digits consists in 50 speakers, ranging from 6 to 14 years old, uttering 77 sequences of digits. Average pitch and formants are estimated for each utterance and PCA dimension reduction is applied. For each discrete VTL warping value, the state sequence likelihood is computed applying eq. 7 with :

1. $\beta = 1$, corresponding to the standard maximum likelihood based VTLN formalism (denoted *ML*)
2. replacing $P(C|Y, M)$ by 1 for $\hat{C} = \max_C P(C|Y, M)$ and 0 otherwise. This comes to select the best warping factor from the long-term features only (denoted *LT*).
3. $\beta = 0.35$, corresponding to the tuning of β for best performance of the hidden mode formalism described in this paper (denoted *HM*).

	reference	ML	LT	HM
Boys	8.8	3.6	3.8	3.2
Girls	11.0	3.4	3.4	3.2
All	9.9	3.5	3.6	3.2

Table 2. Word error rates (in %) for TI-Digits children utterances applying hidden-mode formalism to VTLN. VTLN estimates are obtained on a per-utterance basis.

Table 2 presents the recognition performances. The reference result is obtained with the baseline acoustic model trained on adult speakers with no VTL normalization.

Besides the improved recognition accuracy (8.6% further relative improvement of *HM* with respect to *ML*), this method also yield faster decoding as search paths pertaining to unlikely values of C can be pruned more efficiently if a beam search strategy is used. Indeed, the different search paths are weighted by the likelihood estimate related to the considered value of C .

4. CONCLUSIONS

In the paper, we have shown how the approach of modeling speech signal using multiple (specific, compensated or adapted) models naturally derives from the introduction of additional hidden variables representing long-term stable structures in the speech signal, like the speaker identity or the speaking rate. The formalism that is exposed implies and allows the exploration of features representing both short-term and long-term speech behavior.

This has been evaluated for recognition using vocal tract length (VTL) normalization. Features based on long-term pitch and formant measures, as well as PCA reduction of these have been investigated. Result are conclusive both in terms of ASR accuracy and speed of decoding.

We believe that similar spirited formalisms as well as the kind of feature extraction schemes they imply, should be revisited in more detail. The development of such more complex models is made possible given the current capabilities for making use of larger and more diverse training corpora. Besides improving modeling for ASR, the proposed architecture may also provide estimates of hidden speaking modes, such as the speaker age, the speaking rate, the style, or the accent.

Further research should also focus on studying the independence assumptions made in the proposed formalism.

Acknowledgements

This work has been partly supported by the EU 6th Framework Programme, under contract number IST-2002-002034 (DIVINES project). The views expressed here are those of the authors only. The Community is not liable for any use that may be made of the information contained therein.

5. REFERENCES

- [1] L. Lee and R.C. Rose, "Speaker normalization using efficient frequency warping procedures," in *Proc. of Int. Conf. on Acoustics, Speech, and Signal Processing*, May 1996.
- [2] A. Miguel, E. Lleida, R. Rose, L. Buera, and A. Ortega, "Augmented state space acoustic decoding for modeling local variability in speech," in *Proc. of Interspeech-2005*, Sept. 2005.
- [3] J.J. Odell P.C. Woodland and, V. Valtchev, and S.J. Young, "Large vocabulary continuous speech recognition using htk," in *Proc. of Int. Conf. on Acoustics, Speech, and Signal Processing*, Apr. 1994, pp. 125–128.
- [4] T. Pfau and G. Ruske, "Creating hidden markov models for fast speech," in *Proc. of Int. Conf. on Spoken Language Processing*, Sept. 1998.
- [5] A. Faria and D. Gelbart, "Efficient pitch-based estimation of vtlN warp factors," in *Proc. of Interspeech*, Lisboa, Portugal, Sept. 2005.
- [6] B.E.D. Kingsbury, N. Morgan, and S. Greenberg, "Robust speech recognition using the modulation spectrogram," *Speech Communication*, vol. 25, no. 1-3, August 1998.
- [7] Andreas Tuerk, *the State Based Mixture of Expert HMM with Applications to the Recognition of Spontaneous Speech*, Ph.D. thesis, Cambridge university Engineering Department, sep 2001.
- [8] Stéphane Dupont, Christophe Ris, Laurent Couvreur, and Jean-Marc Boite, "A study of implicit and explicit modeling of coarticulation and pronunciation variation," in *Proc. of Interspeech*, Lisboa, Portugal, Sept. 2005.
- [9] "Divines project web site," <http://www.divines-project.org>, 2004.
- [10] T. A. Stephenson, M. M. Doss, and H. Bourlard, "Speech recognition with auxiliary information," *IEEE Trans. on Speech, Acoustics and Audio Processing*, vol. SAP-12, no. 3, pp. 189–203, 2004.
- [11] Sébastien Poitoux, Olivier Deroo, Christophe Ris, and Stéphane Dupont, "Children speech recognition: a practical implementation," in *submitted to SRIV Workshop - ITRW on Speech Recognition and Intrinsic Variation*, Toulouse, France, May 2006.
- [12] P. Ladefoged and D.E. Broadbent, "Information conveyed by vowels," *The Journal of the Acoustical Society of America*, , no. 29, pp. 98–104, 1957.