

ACOUSTIC SCORES AND SYMBOLIC MISMATCH PENALTIES IN PHONE LATTICES

Louis ten Bosch, Annika Hämmäläinen, Odette Scharenborg, Lou Boves
CLST, Radboud University Nijmegen, Nijmegen, The Netherlands
{L.tenBosch, A.Hamalainen, O.Scharenborg, L.Boves}@let.ru.nl

ABSTRACT

This paper builds on previous work aimed at unraveling the structure of the speech signal using probabilistic representations. The context of this work is a multi-pass speech recognition system in which a phone lattice is created and used as a basis for a lexical decoding pass (search) that allows symbolic mismatches at certain costs. The focus is on the optimization of the costs of the phone insertions, deletions and substitutions that are used in the lexical decoding pass. Two optimization approaches are presented, one related to a multi-pass computational model for human speech recognition, the other based on a decoding that minimizes Bayes' risks. In the final section, the advantages of the two optimization methods are discussed and compared.

1. INTRODUCTION

Currently, there is a growing interest in revisiting multi-pass approaches for automatic speech recognition (ASR) e.g. [2] [5] [11]. In a multi-pass system, a (weighted) phone lattice is often created in the first pass, followed by a lexical search applying additional specialized decoding steps, or using more detailed information, e.g., morphological and domain knowledge. Compared with an integrated search, there are at least two advantages to such an approach. First, a multi-pass approach is useful when spoken keywords are to be detected from a potentially wide range of domains such as meetings, interviews, voicemails, and lectures (cf. [14]). A second advantage is the greater flexibility with which specialized knowledge sources can be brought to bear in subsequent passes, utilizing cross-word triphones, phonotactic restrictions, morphology, long-span syntax, etc. (e.g., [2]).

Weighted phone lattices have shown to be very versatile in a recently developed computational model for human word processing (SpeM). SpeM is a multi-pass decoder in which a phone recognizer in the first pass generates a phone lattice that is used in the subsequent lexical search module. SpeM has been used to successfully model a number of key results from psycho-linguistic experiments [8][9]. In SpeM, mismatches between the phone sequences in the lattice and the phone representations (originating) from the lexicon are dealt with in a more flexible manner than in previous computational models of human auditory word recognition. However, although SpeM does offer more flexibility, all the experiments conducted so far have applied the same penalty for each substitution (and *mutatis mutandis* also for each insertion and deletion) in the lexical decoding. In other words, only three indiscriminate penalties have been used.

In SpeM, the acoustic scores (costs) in the phone lattice computed by the phone recognizer, and the penalties of the phone mismatches (also called 'symbolic mismatches') interact in a complex way. For example, if the mismatch cost is low, the likelihood of associating phone paths in the lattice with a word sequence will be large (since mismatches are cheap), and therefore the probability of decoding the *correct* word sequence might diminish. On the other hand, if the mismatch penalties are high, phone paths must be canonical (and are therefore less likely to have a low cost) to induce a lexical solution; this evidently decreases the likelihood of finding any lexical solution in the phone graph. Therefore, the correct trade-off between the values of 'symbolic' mismatches on the one hand and the acoustic costs in the phone lattice on the other hand is essential for the success of any lexical search pass that takes the phone lattice as input.

This paper focuses on approaches to find an optimal balance between acoustic scores and symbolic mismatch penalties. More specifically, our aim is to investigate how the costs for insertions, deletions, and substitutions affect the likelihood of finding the phone sequence that corresponds to the correct word sequence (as defined by the annotation on word level), and how these optimal costs relate to the acoustic costs in the phone lattice. To that end, we discuss two related optimization approaches. In Section 2, we will deal with the optimization in the context of the SpeM decoding, in which the optimal values for mismatch penalties have been found by a systematic search based on insight in the structure of the search space. In Section 3, we will deal with another, data-driven way to derive optimal mismatch penalties (Minimal Bayesian Risk Decoding). In the final section, we will relate the two approaches and discuss their advantages and disadvantages.

2. LEXICAL PHONE PATHS AND SYMBOLIC MISMATCHES

Our starting point is a phone lattice generated by a free phone loop, guided by a phone bigram. In general, it is often the case that the canonical phone transcription of a word (sequence) is not present in the phone graph, even though a phone lattice may consist of millions of phone paths. For instance, earlier research has shown that the canonical phone transcription of the utterance was not present in the phone graph for 34.9% of a set of 885 phone lattices that were created with a phone bigram [12]. Therefore, phone insertions, deletions, and substitutions must be dealt with to decode utterances in terms of lexical tokens.

In this section, we investigate which conditions will result in the discovery of correct 'lexical' phone paths in phone lattices if symbolic mismatches are allowed. A phone path is 'lexical' if it is a series of phone sequences corresponding to words in the lexicon,

and ‘correct’ if it is made up of those phone sequences that correspond to the orthographic transcription.

The approach consisted of the following steps:

- 1) For each utterance, a phone lattice is created using acoustic models trained on an independent training set.
- 2) A word search algorithm is used to search phone paths associated with sequences of words – allowing symbolic mismatches (phone insertions, deletions, and substitutions) at a specific cost.
- 3) Symbolic mismatch penalties are chosen to optimize the likelihood of the *correct* lexical phone path being the best among all other lexical phone paths through the lattice.

Below, these steps are described in more detail.

2.1. Data and feature extraction

We used a sub-corpus of the Spoken Dutch Corpus, a 9-million-word database comprising 1000 hours of speech annotated on various tiers (e.g. orthographic, prosodic, part-of-speech) [6]. This sub-corpus contains read speech from a Dutch spoken library for the blind. The material comprises word labels as well as manually verified word-level segmentations.

The data in the sub-corpus were divided into three sets: a training set, test set and a development test set (4027, 687 and 687 sentences, respectively). Table I gives an overview of the number of word tokens, speakers, and the amount of speech material per set.

Table I. Datasets used in this study.

	Training	Test	Development	Total
Orthographic word tokens	45,172	7,917	7,507	60,596
Speakers/ Female/Male	125/ 70/55	125/ 70/55	125/ 70/55	125/ 70/55
Duration (hh:mm:ss)	04:51:27	00:51:34	00:48:13	06:31:14

Feature extraction was carried out at a frame rate of 10 ms using a 25-ms Hamming window. A pre-emphasis factor of 0.97 was employed. 12 Mel Frequency Cepstral Coefficients (MFCCs) and log-energy with corresponding first and second order time derivatives were used. Channel normalization was applied by means of CMN over complete recordings (with a mean duration of 3.5 minutes). For training and testing purposes, the data were chunked to grammatical sentences. The feature extraction was performed using HTK [13].

The training corpus was used to create 39 context-independent acoustic models (including 2 different silence models; all models are 3-state left-to-right HMMs with 8 Gaussians per state) on the basis of the lexical phone transcriptions. The lexicon covered all words in the training, test, and development sets, and contained one pronunciation variant per word.

A phone bigram (‘phonotactic model’) was trained on the lexical phone transcriptions of the training corpus. Since leading and trailing silences as well as inter-word silences are annotated on the word level, this method automatically includes bigrams of the form $P(\text{sil}|\phi)$ and $P(\phi|\text{sil})$ (ϕ denoting an arbitrary phone).

2.2. Phone lattice parameter settings

For the construction of the phone lattices, the phone insertion probability and phone-LM factor were tuned using the development set such that the number of phones of the first-best phone path and the number of phones of the canonical phone transcription were equal on average. This was done since mismatches with respect to these lengths will bias the values for phone insertions and deletions. Furthermore, the LM factor was chosen to be as close as possible to 0 (i.e., the decoding is as unbiased as possible). As a result, the insertion log probability and the LM factor were set to -6 and 4, respectively. The beam width was chosen to be large enough to make sure that the time-averaged number of arcs with *different* phone labels was close to 3, i.e., a plausible number of realistic phonetic alternatives is present in the lattice. On average, the resulting lattices had 810 arcs/second, with 12-18 arcs alive per a time slice of 10 ms. The phone paths contained approximately 12.8 phones/second.

2.3. The lexical decoder and the search space

In this research, the search for the lexical phone paths through the phone lattices was based on an FST decoder. This decoder was constructed by interfacing an HTK phone decoder with the AT&T wFST software [4]. The decoding was implemented by a finite state composition of the phone lattice and an FST. This FST was based on the lexical tree, but expanded by including additional arcs with appropriate costs: arcs that accepted any input and wrote out the null label ϵ (modeling phone insertions), arcs that accepted ϵ and write out a phone label (deletions), and arcs that accepted a phone ϕ_2 and wrote out ϕ_1 (substitutions). All insertions shared the same penalty value (idem for deletions and substitutions, resulting in three penalties).

[12] shows that the penalties for symbolic mismatches are to be chosen within certain bounds related to the acoustic scores in the original phone lattice. The structure of the eventual search space is a *union* of lattices, such that each of these lattices is associated with *exactly one* triplet $[I, D, S]$ of non-negative integers I , D and S (representing the number of insertions, deletions and substitutions, respectively, in that particular lattice). Together the parameters I , D and S determine the cost that must be added to the original acoustic cost distribution. The problem of finding optimal symbolic costs is greatly alleviated by restricting the parameters to those regions in the cost space that avoid these distributions to become disjoint.

2.4. Decoding accuracy

We investigated how the costs for insertions, deletions, and substitutions affect the likelihood of finding the phone sequence that corresponds to the correct word sequence (as defined by the orthographic annotation), and how these optimal costs relate to the acoustic costs in the phone lattice. The *decoding accuracy* is defined as the proportion of phone lattices with the following property: after composition with the lexical FST, the correct lexical phone sequence is the cheapest among all lexical phone sequences. This property will guarantee that the lexical search will be able to correctly recognize all words in the entire utterance.

The search is actually three-dimensional, but the contour plot in Figure 1 shows an example of the behavior of the decoding accuracy as a function of the insertion penalty (along the x-axis) and the substitution penalty (along the y-axis). In the figure, the deletion penalty is constant (2.5).

The performance of 0.68 (68 percent of the lattices had the correct word sequence corresponding to the cheapest path) was obtained for substitution, insertion and deletion penalties of about

3.5, 2.4, and 2.5, respectively. In other words, for 68% of the graphs, the correct lexical phone sequence was the cheapest among all lexical phone sequences. For these ‘cheapest sequences’, the proportion of symbolic mismatches compared with the path length depends on the utterance and varies from 18 to 41 percent. For comparison: using a similar technique, [1] reports an average of 26.6 percent phone mismatches on phonetically labeled *manually* transcribed spontaneous speech. The performance difference can, at least partially, be explained by the fact that we only searched for canonical paths in the lattice, whereas it is evident that the actual pronunciation often deviated from the canonical. If the phone recognizer would detect the ‘exact’ sequence of phones in a careful manual transcription, the minimum mismatch rate for read speech would be about 10%.

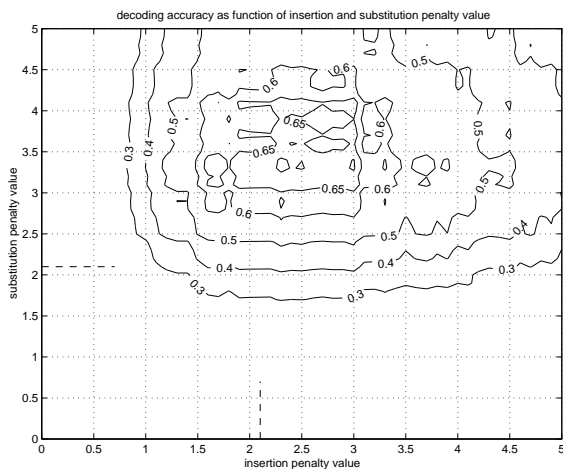


Figure 1. Decoding accuracy as a function of the insertion penalty (along the x-axis) and the substitution penalty (along the y-axis). The horizontal and vertical dashed lines indicate the average acoustic ‘penalty’, i.e. the difference between acoustic score of competing phones.

The optimal symbolic penalties for insertions and deletions are a factor of 1.2 larger than the acoustic mismatch costs, while the substitution penalty is 1.4 to 1.5 times larger than the insertion and deletion penalties. We observe that these ratios are independent of the acoustic costs: since all numerical operations for the total cost along paths are linear, the symbolic penalties scale with the acoustic costs and, thus, the ratios are constant.

3. DATA-DERIVED PHONE-PHONE SUBSTITUTION COSTS

3.1 Minimum Bayesian Risk decoding

In the previous section, the optimal mismatch costs were found by a systematic search that was motivated by the structure of the distribution of path costs. One of the evident drawbacks of such a method is the fact that *all* substitutions are penalized by the same amount, independent of the phonetic or acoustic distance between the source and the target phone. However, the use of many more parameters becomes prohibitive.

Equation 1 provides the mathematical formulation for the optimization of the probability to find the correct phone path in the

phone lattice according to the SpeM decoding framework. The signal X is given, P is the hypothesized lexical path, and Q is a path variable running over the set of all paths available in the phone lattice. The term $-\log(P(X|Q))$ denotes the minus log probability (acoustic score) in the phone lattice, while $d(P, Q)$ denotes the sum of all penalties for symbolic mismatches between the phone sequences P and Q .

$$P_c = \operatorname{argmin}_P \{ \min_Q (-\log(P(X|Q)) + d(P, Q)) \} \quad (1)$$

Shafran & Byrne [10] present a procedure for solving a similar decoding problem. Their approach is based on using a Minimum Bayes Risk (MBR) criterion given by Equation 2:

$$P_c = \operatorname{argmin}_P \sum_Q C(P, Q) P(X|Q) P(Q) \quad (2)$$

in which X , Q and P denote the signal, a path variable (running over all phone sequences in the lattice), and the resulting path, respectively. $C(P, Q)$ denotes the cost of rewriting the path Q into the path P . Their aim is to automatically learn the substitution costs $C(P, Q)$ from data.

It can be seen that Equations 1 and 2 have a similar structure: the \min_Q is replaced by a sum, and $d(P, Q)$ can be interpreted as $C(P, Q)$ (both these terms are basically edit distances). Equation 1 aims at optimizing the probability of finding the correct lexical path in the lattice, while Equation 2 supports the search for optimal penalty values that are valid across all lexical paths in the lattice.

In order to contrast the two approaches, we have defined an iterative scheme using Equation 2 in order to estimate the cost $C(P, Q)$ as follows:

- (1) Decode the data using the current model (acoustic models and current parameter settings that define $C(P, Q)$)
- (2) Compute alignments between the hypothetical transcription from (1) and reference transcription
- (3) Compute the updated $C(P, Q)$ by setting $C(P, Q)$ equal to $-\log(P(P|Q))$

Here, $C(P, Q)$ refers to the weighted edit distance, i.e. the minimum weighted number of modifications to be applied to the phone path Q to obtain the phone path P . The scheme was applied using the same speech data as in Section 2. To increase robustness of the algorithm, we did not introduce new substitution cost parameters for each phone-phone combination – instead, substitution costs were trained for all combinations of five broad phonetic manner classes: plosive (stop), fricative, liquid, nasal, and (semi)vowel. The initial choice for bootstrapping these 25 substitution costs is given in Table IIa. This table is inspired by the optimal substitution penalty found by the SpeM decoding. Table IIb illustrates the result of applying the iteration scheme on all phone lattices of the training set after the third iteration. The third iteration was chosen since, from this iteration on, all matrix entries differ less than 1 percent compared with the values obtained after the third iteration. The evolution of the decoding accuracy (‘Acc’) for the first 5 iterations is presented in Table III. This shows that the Shafran-Byrne scheme is potentially able to outperform the SpeM search due to the feasibility of training more fine-grained mismatch costs – something that was not feasible with the SpeM approach.

Table II shows the penalty values in the same scale as used in the previous section. That means that they can be compared with 2.1, the average acoustic cost for a symbolic mismatch.

3. DISCUSSION AND CONCLUSIONS

The balance between acoustic costs and symbolic mismatches is important for the performance of a multi-pass speech decoding system. In this paper, we discuss two approaches for finding the optimal balance, one in the context of the SpeM decoding and one based on Minimum Bayesian Risk (MBR) decoding. The underlying mathematical formulation of the two methods is very similar. The advantage of the SpeM decoding is that the structure of the search space is known in terms of the three penalty values: the cost distribution is an overlay of smaller distributions that are spaced apart according to the symbolic mismatch values. This structure simplifies the optimisation, because it allows the restriction of the search to specific sub-regions. We have shown that the MBR approach is able to train more refined categories of mismatch costs (by distinguishing more phone classes) methods.

Table II (a, top): Initial substitution cost matrix. (b, bottom): Cost matrix after three iterations.

(a)	Plos	Fric	Liq	Nas	Vowel
Plos	0.0	3.5	3.5	3.5	3.5
Fric	3.5	0.0	3.5	3.5	3.5
Liq	3.5	3.5	0.0	3.5	3.5
Nas	3.5	3.5	3.5	0.0	3.5
Vowel	3.5	3.5	3.5	3.5	0.0
(b)	Plos	Fric	Liq	Nas	Vowel
Plos	0.4	3.2	3.6	4.0	5.1
Fric	3.0	0.6	3.5	3.6	3.9
Liq	3.1	3.5	1.1	2.3	2.8
Nas	3.6	3.9	2.2	0.9	3.2
Vowel	5.2	3.9	2.6	3.4	1.4

Table III Decoding accuracy in percent before the optimization (Iteration 0) and after a number of iteration steps (Iteration 1 to 5) using the MBR optimization scheme.

Iteration	0	1	2	3	4	5
Acc (%)	68	71	71	72	72	72

It is interesting to observe that the costs trained for within-broad-phonetic-class substitutions by the MBR approach are larger than 0. One might expect them to be identically zero. However, these positive values can be explained by the fact that these substitution costs also account for substitutions between non-equal plosives, non-equal nasals (e.g. C(/p/, /q/), C(/m/, /n/) etc..

It is expected that further improvements can be obtained by allowing more fine-grained distinctions in the cost function, e.g. by applying phone-phone dependent substitution costs. Focus of research in the near future will be on the relationship between the acoustic costs, symbolic mismatch penalties, and the decoding of speech in terms of lexical tokens, applied on larger corpora of read speech and on spontaneous speech.

REFERENCES

- [1] Bates, R., and Ostendorf, M. (2002). Modeling pronunciation variation in conversational speech using prosody. Workshop PMLA, Estes Park, USA.
- [2] Demuynck, K., Laureys, T., van Compernelle, D., Van hamme, H. (2003). FLVoR, a flexible architecture for LVCSR. Proceedings *Eurospeech*, Geneva, Switzerland. pp. 1973-1976.
- [3] Frankel, J., Wester, M., King, S. (2004). Articulatory feature recognition using dynamic Bayesian networks. Proc. *INTERSPEECH 2004*, Korea (cdrom).
- [4] Mohri, M., Perreira, F., Riley, M (2003). AT&T FSM Library™, General-Purpose Finite-State Machine Software.
- [5] Ohtsuki, K., Hiroshima, N., Matsunaga, S., Hayashi, Y. (2004). Multi-pass ASR using vocabulary expansion. Proc. *INTERSPEECH 2004*, Korea (cdrom).
- [6] Oostdijk, N. (2002). The design of the Spoken Dutch Corpus. In: Peters, P., Collins, P., Smith A. (Eds) *New Frontiers of Corpus Research* (pp. 105-112). Amsterdam: Rodopi.
- [7] Ristad E.S. and Yianilos, P.N. (1998). A Surficial Pronunciation Model. Proc. *ECSA Workshop on Modeling Pronunciation Variation for ASR*, Rolduc, the Netherlands.
- [8] Scharenborg, O., ten Bosch, L., Boves, L. (2003). ‘Early Recognition’ of Words in Continuous Speech Proc. *ASRU workshop*, St Thomas, US Virgin Islands (cdrom).
- [9] Scharenborg, O., Norris, D., ten Bosch, L., McQueen, J.M., (2005). “How should a speech recognizer work?”. *Cognitive Science*.29(6), 867-918.
- [10] Shafran I. and Byrne, W. (2004). Task-Specific Minimum Bayes-Risk Decoding using Learned Edit Distance. Proc. *INTERSPEECH 2004*, Korea, 1945-1948.
- [11] Tang, M., Seneff, S., and Zue, V. (2003). Two-stage speech recognition using feature-based models: a preliminary study. Proc. *ASRU workshop*, St Thomas, US Virgin Islands (cd).
- [12] Ten Bosch, L., Scharenborg, O. (2005). ASR Decoding in a Computational Model of Human Word Recognition. Proc. *INTERSPEECH 2005*, Lisbon, Portugal , 1241-1244.
- [13] Young, S., Evermann, G., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., and Woodland, P. (2002). *The HTK Book (for HTK Version 3.2.1)*. Engineering Department, Un. Cambridge, UK.
- [14] Yu, P., and Seide, F. (2004). A hybrid-word/phoneme-based approach for improved vocabulary-independent search in spontaneous speech. Proc. *INTERSPEECH 2004*, Korea, 293-296.