

# DEALING WITH PHRASE LEVEL CO-ARTICULATION (PLC) IN SPEECH RECOGNITION: A FIRST APPROACH

Roeland J. F. Ordelman#, Arjan J. van Hessen#, David A. van Leeuwen\*

# University of Twente, Enschede, The Netherlands

\* TNO - Human Factors Research Institute, Soesterberg, The Netherlands

## ABSTRACT

Whereas nowadays within-word co-articulation effects are usually sufficiently dealt with in automatic speech recognition, this is not always the case with phrase level co-articulation effects (PLC). This paper describes a first approach in dealing with phrase level co-articulation by applying these rules on the reference transcripts used for training our recogniser and by adding a set of temporary PLC phones that later on will be mapped on the original phones. In fact we temporarily break down acoustic context into a general and a PLC context. With this method, more robust models could be trained because phones that are confused due to PLC effects like for example /v/-/f/ and /z/-/s/, receive their own models. A first attempt to apply this method is described.

## 1. INTRODUCTION

The DRUID<sup>1</sup> project (Document Retrieval Using Intelligent Disclosure), a collaboration of CTIT<sup>2</sup>/University of Twente, TNO<sup>3</sup> and CWI<sup>4</sup>, aims at the development of tools for the indexing of multimedia content. For the Spoken Document Retrieval (SDR) part of this project, we use ABBOT, the hybrid connectionist-hidden Markov model large vocabulary speech recognition system [1,2] developed for English by Cambridge University, Sheffield University and SoftSound. TNO already participates in the annual English TREC SDR tracks with this system [3], but since the DRUID project focuses on Dutch SDR, we are currently developing a Dutch version of ABBOT.

ABBOT uses a recurrent neural net (RNN) for acoustic modelling and a Markov process for language modelling. Since the RNN is able to capture temporal acoustic context, very good recognition results can be achieved using context-independent phone models. Although language modelling often makes it possible to transform sets

of erroneously recognised phones into well-recognised words, better phone recognition undoubtedly leads to better word recognition.

Our first target was training the phone models in a baseline training, which eventually performed a 33.3% Phone Error Rate (PER) on the test data. Next steps should involve improving acoustic modelling and starting language model training in order to be able to do word recognition. Following on a more detailed description of our methods to improve acoustic modelling in the next sections, this paper reflects our first attempt of improving acoustic modelling by applying phrase level co-articulation rules on the reference transcripts used for training the phone models.

## 2. ACOUSTIC MODEL

### 2.1. Acoustic Training Data

The baseline training material consisted of about 7 hours of speech material of 52 (26 male - 26 female) speakers reading 66 sentences from a newspaper text database, recorded in a noise free room (TNO-NRC-0 database). PLP feature vectors (12<sup>th</sup> order cepstral coefficients derived using perceptual linear prediction and log energy) were presented at the input of the RNN that contained 256 state units. Our phoneset consisted of 44 context-independent phones plus silence.

Obviously, we need far more and also different types of training data to build robust phone models for speaker independent continuous speech recognition in typical SDR tasks, but it is quite an effort to collect large annotated speech corpora for Dutch. Currently we are collecting and annotating speech material from Dutch radio shows and recordings of sessions of parliament.

### 2.2. Annotations

From some of the speech material we are collecting, text auto cues (text to read for newsreader) or annotations (recording and annotation is in special cases a statutory requirement) are available that could reduce at least some of the hard labour. More important, it can provide additional context specific training data for language modelling. Also, CEEFAX documents of the recorded news broadcastings are collected in

<sup>1</sup>

<http://www.seti.cs.utwente.nl/Parlevink/Projects/druid.html>

<sup>2</sup> Centre for Telematics and Information Technology

<sup>3</sup> Institute of Applied Physics, departments Multimedia Technology (Soesterberg) and Human Factors (Delft)

<sup>4</sup> Centre for Mathematics and Computer Science, Amsterdam

order to expand these specific contexts even more which could be useful for the final SDR tasks.

### 2.3. New Phone Set And Transcriptions

Until now, we used the CELEX lexical database [4] for the grapheme to phoneme (G2P) conversion of the annotated text. However, this database contains uncommon, old-fashioned or even incorrect transcriptions so in principle every word has to be checked. In addition, the use of the CELEX database enforces us to adhere to with the choices made by composers of the database. To increase flexibility and have up-to-date and correct transcriptions, we are developing our own G2P tool. This G2P is based on the learning algorithm of Antal van den Bosch [5] and trained on the Van Dale<sup>5</sup> pronunciation dictionary. This dictionary contains less errors and more important a set of up to 200 different phones which gives us the opportunity to get more accurate and flexible transcriptions. For example, the Van Dale dictionary provides the phones /p2/ and /n0/ like in the word 'droppen' (to drop) that is transcribed as /drOp2p@n0/. We may use this full transcription or decide to drop the /p2/ or maybe even the /n0/ if using these special phones turns out to be not of any use. In the CELEX database 'droppen' is transcribed as /drOp@/, leaving out the double /p/ and the very weakly pronounced final /n/.

While we were waiting for additional training data and this new G2P, we have tried to improve acoustic modelling with the available material using co-articulation rules on the phrase level.

### 2.4. Phrase Level Co-articulation

In most G2P algorithms (dictionary look-up as well as rule based), within-word co-articulation effects are usually sufficiently dealt with. However, with co-articulation effects on the phrase level (PLC) this is not always the case. Yet, the use of G2P algorithms that do PLC rules for training as well as decoding, should be able to improve recognition performance.

During decoding, phrase level co-articulation can be modelled as proposed in [6] for example. In this approach words are joined together under certain criteria, thus forming 'multiwords' that are added to a lexicon and phonetically transcribed according to phonological rules.

Also on the level of acoustic modelling during training, the possibility to deal with PLC effects should be incorporated. When phrases are solely transcribed on a word-by-word basis, reference transcriptions can contain wrong phones or phones that should not be there due to these crossword co-articulation effects. Consequently, the recogniser is

trained absent or wrong phones. This obviously results in less than optimal phone models.

With this in mind we figured that although the recurrent neural net of the ABBOT system can very well deal with acoustic context, providing the RNN with better reference transcripts during training should improve training performance and therefore also recognition performance. Especially closely related phones (further on called 'confusion phones') like /n/ - /m/, /f/ - /v/, /s/ - /z/ and to a lesser degree /d/ - /t/ and /p/ - /b/ that nearly all mainly differ in the voicing feature, could benefit from applying PLC rules since in Dutch progressive and regressive voice assimilation are frequently occurring phenomena.

Phone	%Error	Conf phone	%Conf
/z/	54	/s/ (+)	66
/f/	54	/v/ (+)	47
/v/	14	/f/	27
/p/	13	/b/	26
/b/	29	/p/ (+)	26
/d/	40	/t/	12
/s/	11	/z/	9
/t/	22	/d/	6
/n/	33	/m/	3

**Table I:** Individual phone error rates with the corresponding typical confusion phones and their frequency percentage. A "(+)" indicates that this confusion caused most of the errors.

In Table I individual phone error rates are shown with the corresponding typical confusion phones. The column "%Conf" gives the percentage of error that was caused by the confusion phone. A "(+)" indicates that this confusion caused most of the errors. In all other cases, most of the errors were due to deletions, the confusion phones following shortly after.

In order to apply PLC rules on the acoustic modelling level we selected three phonological rules of Dutch that are frequently applied within words. Generally these rules also apply on the phrase level provided that there is no pause between two succeeding words that prevents such a co-articulation process:

- Regressive voicing/devoicing of plosives and fricatives

$\alpha$ voiced (fricative/plosive)  $\rightarrow$   
 $-\alpha$ voiced | \_#  $-\alpha$ voiced (plosive)

Examples:

(hij) gaf dit (aan mij)  $\rightarrow$  /xAv dIt/ (He gave this to me)

(hij) las de (boeken)  $\rightarrow$  /lAz d@/ (He read the books)

<sup>5</sup> Van Dale – Dutch Dictionary Publisher:  
<http://www.vandale.nl/>

- Progressive devoicing of fricatives

voiced (fricative) →

-voiced | (fricative/plosive) #\_

Examples:

(Ik) beloof ze → /b@lo:f s@/ (I believe them)

(Ik) liep voorbij → /li:p fo:t/ (I walked by)

- Nasal adaptation

/n/ → /m/ | \_# /p,b,m/

Examples:

(Hij woont) in Belgie → /Im bElgi:j@/ (He lives in Belgium)

(De) man praat → /mAm pra:t/ (The man talks)

We did not select the Dutch deletion rules although they are frequently applied, especially in conversational speech. The reason we did not select them is a practical one: as these rules 'eat away' final/initial phones we could end up with non-existing words as for example in the Dutch phrase 'in Nederland' (in the Netherlands). In Dutch the deletion rule exists that of two equal and adjacent phones, one of them is deleted. If this rule was applied on phrase level, we would end up with the phone sequence /Ine:d@rIAnt/ leaving us with a problem when trying to map this phone sequence to the two words 'in' and 'Nederland'. In addition, strange ambiguities might occur. An example is the Dutch phrase '(er is) *nog geen* enkel (bericht)' (there is **not one** single message yet) that would be transcribed as /nOxe:n/ (there is **only one** message left) applying this same deletion rule.

One can argue that the same problem appears while applying the PLC rules mentioned above. However, as will be explained below, we avoided this by using descriptions of the new phones in such a way that the original phone could be restored.

Yet, deletions are a problem in continuous speech recognition. Although it is not very useful trying to recognise phones that are not really present, and adapting a system to train without those 'ghost' phones is not that much of a problem either, the question however remains how to deal with the decoding problem. A solution might be applying PLC rules on-line, during decoding. We planned to investigate this possibility in the future but for now we take the point that deletions should not be addressed on the acoustic modelling level but rather on the decoding level.

### 3. METHOD

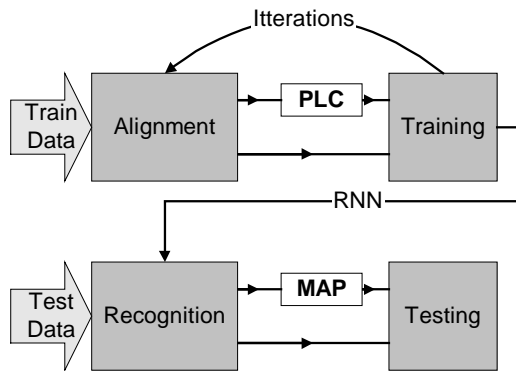
In order to apply PLC rules during acoustic modelling we plugged them into our baseline training set-up. In our standard training procedure acoustic feature representations of the training data are passed to the current RNN, being a result of a previous training. This network produces phone posterior probabilities for each frame of the data. A phone label is assigned to these frames using a Viterbi alignment. Based on this phone/frame alignment the RNN is trained. With this new RNN we can update the phone/frame alignment and train the RNN again with the improved alignment. The training procedure is fully described in [1].

We plugged in the PLC rules after the alignment procedure and applied them to the updated phone alignment by simply replacing phones that would be effected by phrase level co-articulation with new phones (see Fig. 1). The phone /v/ that ought to be devoiced due to a preceding /p/ (rule 2) for example, was altered to the phone /vu/: a /v- 'unvoiced/'. In Table II the complete mapping list used is shown. Avoiding the creation of new phones by just altering this /v/ in a /f/ was not an option, because in doing such, we would loose the link to the words in the lexicon.

By applying PLC rules this way, we can prevent that a phone that is influenced by PLC effects 'contaminates' the training of that particular phone. We can therefor build more robust models. But then we also take away some of the context sensitivity of the phone by redirecting some of its context. This loss can be taken care off by the PLC phones that are mapped back to the original phone later on. In other words, we break down acoustic context into a general context that we use to train the original phone, and a PLC context to train the PLC phone and merge these contexts later on during decoding.

Phone	Mapping	Phone	Mapping
buv	b	Pv	P
duv	d	Tv	T
gv	g	Kv	K
Zuv	Z	Sv	S
zuv	z	Sv	S
vuv	v	Fv	F
Guv	G	Xv	X
nm	n		

**Table II:** New phones and mapping



**Fig I:** PLC Training procedure

With the 'enriched' alignment the RNN was trained. After a recognition pass with this RNN, the new phones of the test data were mapped onto the original phones in order to be able to compare the output of the neural net (phone recognition) with its input (reference transcript). Figure I shows our PLC training procedure. We ran two tests, one with our baseline RNN (without PLC rules) and one with the new RNN (with PLC rules).

## 4. RESULTS

In Table III the phone error rates of the two different training methods are shown. We have listed only the confusion phones because these are the phones we are especially interested in. IPER refers to Individual Phone Error Rate (errors of individual phones divided by frequency of occurrence). The PER is the overall Phone Error Rate (including all phones). The three columns at the right contain the amount of errors for a particular phone that this phone was confused by a phone in the second right column ('Conf Phn'), next to the total number of errors for that phone.

For example on the first line, the phone /b/ (after baseline training) has a individual phone error rate of 29.35 % which improves (after PLC training) to a 27.65% IPER. The phone /b/ was 22 out of 86 times wrongly recognised as /p/ after baseline training and 22 out of 81 times after PLC training.

It looks as if the use of PLC rules has made a small difference. Although overall the PLC training hardly performs better, the individual phone error rates of the phones /b/, /d/, /z/, /G/, /S/ and /f/ have decreased. The right side of Table III, shows that in three out of six times this improved performance can not be caused by a decrease of confusions. Some phones that overall do better after the PLC training, showed an increase of their number of confusions (/S/, /z/) or no change at all (/b/) in comparison to the baseline training performance. The better performance of /f/, /G/ and /d/ after PLC training does run parallel with a decrease of confusions but the total amount of errors for these phones decreases too. Furthermore, there is a

downswing of performance and an increase of confusion errors of the phone /t/ and in particular the phone /v/.

Phn	Bsln IPER (%)	PLC IPER (%)	Cnf Phn	Bsln High. Conf.	PLC High. Conf.
b	29.35	27.65	p	22/86	22/81
d	40.27	37.14	t	42/360	39/332
g	100	100			
Z	100	100			
z	54.26	53.10	s	92/140	96/137
v	14	32.15	f	19/71	75/163
G	59.01	50.93	x	48/95	41/82
n	32.99	33.65	m	14/446	32/455
p	12.95	12.35	b	11/43	10/41
t	21.59	21.80	d	19/307	30/332
k	12.48	13.23	T	12/66	11/70
S	72.73	69.7	S	9/24	12/23
s	10.99	10.51	Z	18/91	14/87
f	54.02	46.55	v	44/94	37/81
x	23.08	25.11	G	23/102	29/111
PER	33.34	32.91			

**Table III:** Phone error rates of the two different training methods. In the first column the phone we are interested in, followed by two columns with individual phone error rates. The columns on the right contain confusion percentages.

## 5. CONCLUSION

Since the amount of training data for the baseline training has been relatively small we were prepared for only a small effect of applying PLC rules, especially because of our approach of introducing new phones with a low frequency of occurrence. The small amount of data forces us to be careful drawing conclusions, but it seems we have succeeded in building more robust phone models of phones in general context. By removing PLC context from the training of these models, the phones are slightly better recognised which is reflected by the decreased individual phone error rates in a majority of the cases. On the other hand, we see some cases that perform worse after PLC training or have a larger amount of confusions than they had after baseline training. This is most probably due to the fact that the PLC phones are not trained very well yet.

## 6. DISCUSSION

Much more training data is needed (and coming up) to train more extensively the acoustic models of our recogniser. With this material we intend to explore the usefulness of applying phrase level co-articulation rules on the acoustic modelling level further.

In order to be able to apply deletion rules on the phrase level as well, we will investigate the possibility of applying PLC rules both during

training and decoding as discussed earlier (2.4) In such an approach phones will first be altered or deleted in the training transcripts so that PLC context will be removed. The phone models are then trained without the PLC context. When during decoding a PLC context is seen, a particular phone can be altered or inserted according to (reverse) PLC rules.

An advantage of this approach is that there is no need to introduce new PLC phones that are hard to train because of their infrequent occurrence. A disadvantage however is that this method is fairly crude as it neglects the intermediate status of phones that are influenced by (phrase-level) co-articulation effects.

## REFERENCES

- [1] Tony Robinson, Mike Hochberg and Steve Renals, *The use of recurrent neural networks in continuous speech recognition*, <http://svrwww.eng.cam.ac.uk/~ajr/rnn4csr94/rnn4csr94.html>
- [2] Morgan, N., Bourlard, H., *Connectionist Speech Recognition: A Hybrid Approach*. Kluwer Academic Publishers, 1994.
- [3] Wessel Kraaij, Joop van Gent, Rudie Ekkelenkamp and David van Leeuwen, "Phoneme based Spoken Document Retrieval", *Proceedings of TREC-7*, NIST, 1998
- [4] <http://www.kun.nl/celex/>
- [5] Antal P. J. van den Bosch, *Learning to pronounce written words, A study in inductive language learning*. Thesis, University of Maastricht, 1997
- [6] Mirjam Wester, Judith M. Kessens and Helmer Strik, "Improving the performance of a Dutch CSR by modelling pronunciation variation", *Proc. of the Workshop Modelling Pronunciation Variation for Automatic Speech Recognition*, Kerkrade, 145-150, 1998