IMPROVING RECOGNITION PERFORMANCE USING CO-ARTICULATION RULES ON THE PHRASE LEVEL: 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

This paper describes a first approach to improve recognition performance of our hybrid large vocabulary continuous speech recogniser for Dutch by using co-articulation rules on the phrase level. By applying these rules on the reference transcripts used for training the recogniser and by adding a set of special temporary phones that later on will be mapped on the original phones, more robust models of phones that are typically confused a lot in speech recognition like /v/-/f/ and /s/-/z/, could be trained.

1. INTRODUCTION

The DRUID project (Document Retrieval Using Intelligent Disclosure), a collaboration of CTIT/University of Twente, TNO[1] and the CWI[2], 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 English TREC SDR track with this system [3], but since the DRUID[3] 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 is often able 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, we report about our first approach of improving acoustic modelling by applying phrase level co-articulation rules on the reference transcripts used for training the recogniser.

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). Obviously, we need far more and also different types of training data to train robust phone models for speech recognition in typical Spoken Document Retrieval 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 these, 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 order to expand these specific contexts even more. This could be useful for the final Spoken Document Retrieval task.

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 that in principle every word has to be checked. In addition, the use of the CELEX database enforces us to go 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[4] pronunciation dictionary. This dictionary contains less errors and more important a set of up to 200 different phones which gives us the opportunity to be more accurate and flexible. For example, the Van Dale dictionary provides the phones /p2/ and /n0/ like

[1] Institute of Applied Physics, departements Multimedia Technology and Human Factors

[2] Centrum voor Wiskunde en Informatica (Centre for Mathematics and Computer Science), Amsterdam: http://www.cwi.nl

[3] http://www.seti.cs.utwente.nl/Parlevink/Projects/ druid.html

[4] Van Dale – Dutch Dictionary Publisher: http://www.vandale.nl/

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 even the /n0/ if it turns out to be too not of any use.

While we are 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, phrases transcribed solely on a word-by-word basis, can contain wrong phones or phones that should not be there due to crossword co-articulation effects. Consequently, the recogniser is trained on phones that may be absent or wrong, which obviously results in less than optimal phone models. Continuous Speech Recognition should therefor also incorporate the ability to deal with phrase-level co-articulation (PLC) effects.

With this in mind we figured that although the recurrent neural net of the ABBOT system can deal very well with acoustic context, providing the neural net with 'enriched' reference transcripts during training should improve training performance and therefor also recognition performance. Especially closely related phone pairs (that we will call 'confusion phones' further on) 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. In Table I the 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.

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.

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

αvoiced (fricative/plosive) →
-αvoiced | _# -αvoiced (plosive)

Examples:
(hij) gaf dit (aan mij) → /xA**v d**It/ (He gave this to me)
(hij) las de (boeken) → /lA**z d**@/ (He read the books)

- 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 f**o:r/ (I walked by)

- Nasal adaptation

/n/ → /m/ | _# /p,b,m/

Examples:
(Hij woont) in Belgie → /I**m b**Elgi:j@/ (He lives in Belgium)
(De) man praat → /mA**m p**ra:t/ (The man talks)

We did not select the Dutch deletion rules although they are frequently applied, especially in conversational speech. The reason we didn't 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 phase 'in Nederland' (in the Netherlands). If on the phrase level the deletion rule was applied that from two equal and succeeding phones, one of them is deleted, we would end up with the phone sequence /Ine:d@rlAnt/ leaving us with a problem when trying to map this phone sequence to the two words 'in' and 'Nederland'. In addition, strange ambiguities may 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 this same problem appears while applying the rules mentioned above. However, as 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. They should however not be addressed on the acoustic modelling level but rather on the language modelling level or by adding multiple transcriptions to the lexicon as proposed in [6].

3. METHOD

In order to study the effect of applying PLC rules during acoustic modelling we plugged 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. Using a Viterbi alignment these frames are assigned a phone label and 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 after the alignment procedure and applied PLC rules on the updated alignment by simply replacing phones that would be effected by phrase level co-articulation with new phones (see Fig. I). 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 doing such, we would cut off the road to the words in the lexicon.

Phone	Mapping	Phone	Mapping
b*uv*	b	P*v*	P
d*uv*	d	T*v*	T
g*v*	g	K*v*	K
Z*uv*	Z	S*v*	S
z*uv*	z	S*v*	S
v*uv*	v	F*v*	F
G*uv*	G	X*v*	X
n*m*	n		

Table II: New phones and mapping

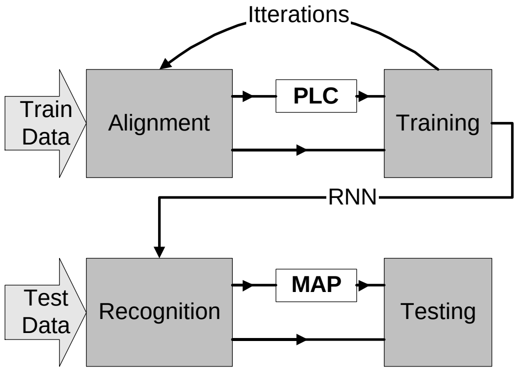

Fig I: PLC Training procedure

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.

4. RESULTS

In Table III the phone error rates of the two different training methods are shown. IPER refers to Individual Phone Error Rate (errors of individual phones divided by frequency of occurrence). The Total 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. Phone'), 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. We are interested in the performance of these confusion phones in particular since an effect of applying PLC rules should show up here.

Phn	Bsln	PLC	Cnf	Bsln	PLC
	IPER	IPER	Phn	High.	High.
	(%)	(%)		Conf.	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
Tot %	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.

A first glance at the columns at the left of Table III could give the impression that the use of PLC rules has made a small difference since the PLC training performs overall slightly better and has also decreased the individual phone error rates of the phones /b/, /d/, /z/, /G/, /S/ and /f/. The right side of Table III however shows that in three out of six times this improved performance can not be caused by applying PLC rules. PLC rules should decrease the amount of confusions but some phones that overall do better after the PLC training, saw their number of confusions increase (/S/, /z/) or stay the same (/b/) by comparison with the baseline training performance. The better performance of /f/, /G/ and /d/ after PLC training does run parallel with a decrease of confusions but whether applying PLC rules has something to do with this is questionable, since the total amount of errors for these phones decreased also. Furthermore, there is a downswing of performance and an increase of confusion errors of the phones /v/ and /t/.

Reference	Recognised	Eligible for PLC
p	b	1
v	f	5
z	s	31

Table IV: Occurrences of phone sequences eligible for applying PLC rules

To get a better picture of the phrase level co-articulation processes that could have taken place in the test data we counted all occurrences of phone sequences that are eligible for applying PLC rules (Table IV). Only a very small amount of phones appeared to have a neighbouring phone that would force one of our PLC rules to be employed. Apparently we have been trying to model something that was not there at all.

5. CONCLUSION

Of all confusion phones only the phone /z/ and /v/ could benefit from applying co-articulation rules on the phrase level. So for most of the phones under investigation, there was no effect of applying PLC rules to be found! As in the case of the phone /z/, eventually the only phone that could benefit form PLC training, there was no effect on performance, we conclude that lack of results are due to the PLC training itself. Evidently the amount of training data was too small to do a significant PLC training with a lot of new phones.

6. DISCUSSION

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. We did not expect however our chosen phrase level co-articulation rules to be that low frequent. We could have known if we checked this beforehand, but we did not. We still assume though, that applying PLC rules can make a difference. Since much more training data is needed (and coming up) to train more extensively the acoustic models of our recogniser, we intend to explore the usefulness of applying phrase level co-articulation rules on the acoustic modelling level further.

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", P*roc. of the Workshop Modelling Pronunciation Variation for Automatic Speech Recognition, Kerkrade*, 145-150, 1998