EXPLOITING PHONETIC AND PHONOLOGICAL SIMILARITIES AS A FIRST STEP FOR ROBUST SPEECH RECOGNITION

Julie Mauclair, Daniel Aioanei, Julie Carson-Berndsen

School of Computer Science and Informatics, University College Dublin
Belfield, Dublin 4, Ireland
email:{julie.mauclair, daniel.aioanei, julie.berndsen}@ucd.ie
web: http://muster.ucd.ie

ABSTRACT

This paper presents two speech recognition systems which use the notion of phonetic and phonological similarity to improve the robustness of phoneme recognition. The first recognition system, YASPER, uses phonetic feature extraction engines to identify phonemes based on overlap relations between phonetic features. The second system uses the CMU Sphinx 3.7 decoder based on statistical context-dependent phone models. Experiments have been carried out on the TIMIT corpus which show improvements in phoneme error rate when a projection set constructed with respect to phonetic and phonological similarity is used. It is envisaged that in future, the two systems will provide alternative parallel streams of hypotheses for each interval of the speech signal and will work together as experts in the phoneme recognition process.

1. INTRODUCTION

One of the key challenges for a speech recognition system is construct acoustic models which correctly estimate a sub-word unit or phonetic class label for a specific time interval. The information required by such models is influenced by temporal phenomena such as co-articulation (overlap of properties) and is constrained by phonotactic restrictions (precedence relations between properties). Statistical context-dependent phone models are often used to deal with such temporal phenomena. Another approach is to use units with finer granularity than the phone, namely phonetic features, which are then constrained by rules to identify the phones, the syllables and ultimately the words characterised by these features. This paper presents a comparison of the outputs of these two approaches and demonstrates how the notion of phonetic and phonological similarity can be incorporated into both. The underlying motivation for this line of research is to develop a speech recognition system which uses multiple resources and integrates symbolic and statistical information in an efficient an principled manner. It is envisaged that the systems described below will, in the future, form part of an integrated multi-tiered package for phoneme recognition.

Combining outputs from different systems has proved quite successful in speech recognition. The ROVER technique [4] uses a voting process to create confusion networks from the output of several ASR systems. Based on the outcome of the experiments described in this paper, we propose to investigate how to combine two systems at the phoneme level.

One approach to speech recognition which uses phonetic features is the Time Map model [3]. Here an event logic and phonetic and phonological constraints are employed to address the problem of co-articulation. In this case, statistical methods are used at a lower level to identify phonetic features rather than words, syllables or phones. The implementation of the Time Map model which has been used in the experiments described below is YASPER [2, 1]. The second phoneme recognition system has been built using the CMU Sphinx 3.7 decoder [12]. Given the output of each of the systems, projection sets defined with respect to varying criteria are applied and the change in *phoneme error rate* (PER) is measured.

The approach described below is similar to the phoneme classifier presented in [8] in which arrangements of experts based on *Broad Phonetic Group* (BPG) are proposed. The notion of BPG is based on [6] where it was shown that approximately 80% of misclassified frames are those for which the right phoneme is in the same BPG. The BPGs in [8] are defined according to a confusability matrix. A new phoneme classifier is proposed consisting of modular arrangements of experts, with one expert assigned to each BPG and focused on discriminating between phonemes within that BPG. The result in PER achieved by that system on the TIMIT corpus [5] is 26.4%. Elsewhere in the literature, [13, 14, 9], for phoneme recognition experiments using the TIMIT corpus, PER has ranged between 47.3% and 23.4%.

The next two sections of this paper present the phoneme recognition systems used in the experiments. The experiments using the TIMIT corpus are then described in section 4. The results are discussed in section 5 and some conclusions are drawn which provide avenues for future work.

2. PHONEME RECOGNITION USING YASPER

YASPER [2] uses the event logic of Time Map Phonology [3] to analyse phonetic events realised as acoustic-articulatory features detected by a set of parallel feature extractors. The goal of the system is to identify the sounds produced by these phonetic events and to output them efficiently as units (in this case phonemes). The input to the system thus is a set of streams (or tiers) of acoustic-articulatory features as they are extracted from the speech signal.

The output of the feature extraction engines (based on both HMM and SVM) is processed by an interval manager which deals with co-articulation and uses feature implication rules to interpret underspecified and noisy input. The multilinear representation of features is parsed to identify intervals

This material is based upon works supported by the Science Foundation Ireland under Grant No. 07/CE/I1142. The opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of Science Foundation Ireland.

of overlapping features. The intervals are then analysed and validated using feature implication rules. The feature implication rules used by the systems have been derived from a feature hierarchy, which captures the logical dependencies between the features in a phoneme-to-feature mapping used by the system [7]. Once the interval has been validated, the phoneme or the set of phonemes (in the case of underspecified input), characterised by the features in the current interval, is identified. These processes are described in more detail in [1].

3. PHONEME RECOGNITION USING CMU SPHINX

For the experiment presented in this paper, the most recent version of the CMU Sphinx decoder, Sphinx 3.7 [12] was used. This decoder uses continuous and context-dependent HMM acoustic models. However, since it does not have a dedicated phoneme recognition process, all parts of the decoder were trained with phonemes instead of words which are otherwise typically used. The lexicon is thus composed of phonemes rather than words.

The acoustic models were trained on the TIMIT training set using SphinxTrain. The training set consists of approximately 3.1 hours of speech. The phonemes were all modelled with a 3 state, left-to-right HMM with no skip state and 13 MFCCs. The acoustic models were trained with 8000 senones and 8 Gaussians per state; these parameters were tuned based on a development set.

The trigam language model was trained on the TIMIT time-aligned phonetic transcriptions provided for the training corpus thanks to the CMU Statistical Language Modeling Toolkit [11].

4. EXPERIMENTS

The two ASR systems have been tested at a phoneme level on the test set provided by TIMIT. This test set consists of 1680 sentences and 58191 instances of phonemes for the phoneme transcription provided. The phoneme error rate on TIMIT found in the literature ranges between 47.3% and 23.4% [13, 14, 9]; 23.4% is thus taken as a baseline comparison value for the experiments below.

4.1 Set of phonemes

The set of phonemes used consists of 44 phonemes: aa, ae, ah, ao, aw, ax, ay, b, ch, d, dh, dx, eh, er, ey, f, g, hh, ix, iy, jh, k, l, m, n, ng, ow, oy, p, q, r, s, sh, t, th, uh, uw, ux, v, w, y, z, zh and sil.

The CMU Sphinx-based phoneme recognition system achieves 36.2% in PER on the test set based on the most probable (*1st best*) phoneme sequence.

4.2 Evaluation metrics

In the experiments described below, the scoring is done using the NIST *sclite*, scoring and evaluation tool [10]. Sclite compares the output hypothesis produced by the speech recogniser with a reference. Sclite uses a dynamic programming algorithm to minimise a distance function between pairs of labels from the hypothesis and the reference. In the experiments below which use various projection sets, the hypothesis will consist of a sequence of *sets* of phonemes rather than a sequence of single phonemes. Every phoneme is substituted with all phones in the corresponding class. The reference will still be a sequence of phonemes. During processing a set as an hypothesis, sclite compares all the phonemes that are contained in that particular set with the reference phoneme. An error is counted if the correct phone is not in the set. It is important to note that results are presented in terms of PER since scoring is with respect to phonemes and not in terms of classes or groupings of phonemes.

4.3 Projection sets

To evaluate the use of phonetic and phonological similarity in the phoneme recognition process, a sequence of sets of phonemes can be used instead of a sequence of only the most probable phonemes. Such sets are termed *projection sets* and can be constructed in various ways. In the following subsections, different projection sets are defined and evaluated.

4.3.1 YASPER sets

In [1] , YASPER is evaluated with the output of different combinations of feature extraction engines. Currently, the best results are based on the feature extraction engines for the features *vocalic*, *palatal*, and *voiced* . Thus, the outputs of YASPER correspond to sets of phonemes for each extractor engine : { ch / f / hh / k / p / q / sh / th / sil / s / t } { b / dh / g / jh / m / ng / r / v / w / zh / d / dx / n / z / l / y} { ax / er / aa / ae / ah / ao / aw / ay / eh / ey / ix / iy / ow / oy / uh / uw / ux }. In the table below, the configuration of YASPER with the feature extraction engines for these features will be referred to as VPV (Voiced-Palatal-Vocalic).

The CMU Sphinx-based system provides the most probable phoneme for the current interval. To compare the results given by Sphinx with the output of YASPER, each phoneme in the most probable phoneme sequence is projected to the relevant VPV set. For example, if the phoneme /b/ is recognized, the projection set { b / dh / g / jh / m / ng / r / v / w / zh / d / dx / n / z / l / y} is used in the evaluation.

The two recognition systems results are given in table 2. The remarkable improvement in phoneme error rate for the Sphinx VPV configuration can be put down to the fact that each phoneme in the most probable set is projected onto approximately one third of the full phoneme set and thus results, not unexpectedly, in an increase in recall at the expense of precision. This type of projection set is supported only weakly by a principled phonetic motivation in terms of grouping sounds which have something in common. However, the grouping is too broad and does not exhibit enough distinctive power.

Recognition System	**Sphinx 1-best**	**YASPER VPV**	**Sphinx VPV**
PER	**36.2%**	**31%**	**19.6%**

Table 2: Results at phoneme level for the two phoneme recognition systems with the VPV set

4.3.2 Natural Classes sets

As a second experiment, we evaluate the results projecting the most probable phoneme to a set of *natural classes*. Natural classes represent phonological similarity in that they de-

Phoneme : set of phonetically similar phonemes/**extended set of phonetically similar phonemes**

aa : aa / ah / ao / **vowels**	ux : ux / uh / ix / **vowels**	ng : ng / n / m / **k / g / q / ch / jh / hh / q**
ae : ae / ax / ah / **vowels**	er : er / r / aa / **aa / ae / ah / ao / aw / ay / eh / ey**	p : p / sil / b / f / **g / q / t / d / k / v / m / w**
ah : ah / ae / ax / **vowels**	b : b / sil / p / v / **g / q / t / d / k / f / m / w**	q : q / sil / **g / t / d / k / b / p / ch / jh / hh**
ao : ao / ah / aw / **vowels**	ch : ch / jh / k / **th / dh / sh / zh / q**	r : r / er / l / **w / y / n / t / d / dx / s / z**
aw : aw / aa / ao / **vowels**	d : d / sil / dx / t / **th / dh / b / p / f / v / n**	s : s / z / sh/ **v / f / th / dh / n / t**
ax : ax / ae / ah / **vowels**	dh : dh / th / d / dx / **f / v / s / z / sh / zh / ch / t**	sh : sh / s / zh / **z / th / dh / ch / jh / n / t**
ay : ay / ey / oy / **vowels**	dx : dx / d / t / **th / dh / b / p / f / v / n / sil**	t : t / sil / d / n / **dx / th / dh / b / p / f / v**
eh : eh / ah / ix / **vowels**	f : f / v / s **/ z / sh / th / dh / p / b / m / w**	th : th / dh / t / **f / v / s / z / sh / ch / d / dx**
ey : ey / ay / iy / **vowels**	g : g / sil / k / d / **q / t / p / b / ng / ch / jh**	v : v / f / z / **s / sh / th / dh / p / b / m / w**
ix : ix / ah / eh / **vowels**	hh : hh / jh / ch / **sh / q / k**	w : w / y / l / **b / p / f / v / m**
iy : iy / ey / ix / **vowels**	jh : jh / ch / g / **th / dh / sh / zh / k / q**	y : y / w / l / **sh / zh / jh / ch / k / er / r**
ow : ow / oy / ao / **vowels**	k : k / sil / g / t / **q / d / p / b / ng / ch / jh**	z : z / s / zh / **v / f / sh / th / dh / n / t**
oy : oy / aw / ao / **vowels**	l : l / w / r / **y / er / n / t / d / dx / s / z**	zh : zh / z / sh / **s / t / dh / ch / jh / n / t**
uh : uh / uw / ux / **vowels**	m : m / n / ng / **p / b / f / v / w**	sil : sil / t / k / **p / q**
uw : uw / uh / ax / **vowels**	n : n / m / ng / **t / d / s / z / r / l / dx / dh / th / sh / zh**	

Table 1: Associations of a phoneme with its phonetic similarity set **and extended set**

fine classes of sounds which have language-specific distributional characteristics in common. The projection sets based on natural classes used in the experiment in this paper are the stops - { k / p / q / b / g / d / dx / t / sil}, vowels - { ax / aa / ae / ah / ao / aw / ay / eh / ey / ix / iy / ow / oy / uh / uw / ux }, fricatives - { jh / ch / f / hh / sh / th / s / v / zh / z / dh}, nasals - { m / ng /n}, and approximants - {er / w / l / r / y}.

The results of this experiment are summarized in table 3. The projection sets defined with respect to natural classes provide a more principled way to infer alternative phoneme hypotheses from the most probable phoneme sequence. This is still at the expense of precision, however.

Recognition System	Sphinx 1-best	YASPER VPV	**Sphinx natural classes**
PER	36.2%	31%	**19.3%**

Table 3: Results at phoneme level for the two phoneme recognition systems with the natural classes projection sets

4.3.3 Phonetic similarity (PS) sets

For this experiment, projection sets were constructed with respect to a notion of phonetic similarity which describes the phonetic neighbourhood of a particular sound in terms of its acoustic and articulatory properties. The projection sets are defined in table 1 and distinguish between immediate and near neighbours (the extended set highlighted in bold). Two experiments are carried out, one with only projection sets consisting of only the two closest neighbours and one with the extended projection set.

Although the extended PS perform better than the PS sets, in order to improve precision, smaller sets are preferable. It seems plausible that context information will provide useful information for the selection of the most appropriate projection set for any specific phoneme hypothesis. For this reason, context information is investigated in the next section.

Recognition System	Sphinx 1-best	YASPER VPV	**Sphinx PS**	**Sphinx PS extended**
PER	36.2%	31%	**27.4%**	**17.9%**

Table 4: Results at phoneme level for the two phoneme recognition systems with the phonetic similarity projection sets

4.3.4 Contextual factors

In order to investigate the value of contextual information for the definition of more appropriate PS sets, as a first step the deletions, susbtitutions, and insertions errors of the Sphinx-based phoneme recognition system were analysed. For this experiment, the context is restricted to the immediately preceding and following context of the phoneme in question and does not at this point consider larger linguistic units such as syllables, words or phrases although this has been addressed elsewhere using phonotactic models [1].

Firstly, the phoneme recognition system was evaluated on a development set. This provided a list of every error made by the Sphinx-based system in a particular context (the phoneme to the left and to the right). The test set was then used with the list of errors from the development set to define projection sets based on errors encountered. This is closely related to the notion of confusability but includes specific reference to context. Figure 1 depicts an example of the type of projection sets obtained in this experiment. When there was a substitution of the phoneme p_1 by the phoneme p_2 with the left and right context p_l and p_r in the development list, the phoneme p_2 is then associated with all possible substituted phonemes p_1 in a set $S_{p_l p_r}$ for that context. So, when the pattern $p_l\ p_2\ p_r$ appears in the result provided by the recognizer, it is replaced by $p_l\ S_{p_l p_r}\ p_r$ for the evaluation. For a deletion, the sequence $p_l\ p_r$ has been recognized instead of $p_l\ p_d\ p_r$. All occurences of p_d are then grouped in $S_{p_l p_r}$ and the sequence $p_l\ p_r$ is replaced by $p_l\ S_{p_l p_r}\ p_r$. When an insertion occurs, the recognizer provided $p_l\ p_i\ p_r$ instead of $p_l\ p_r$. The symbol @ is added to $S_{p_l p_r}$ to allow $p_l\ p_r$ to directly

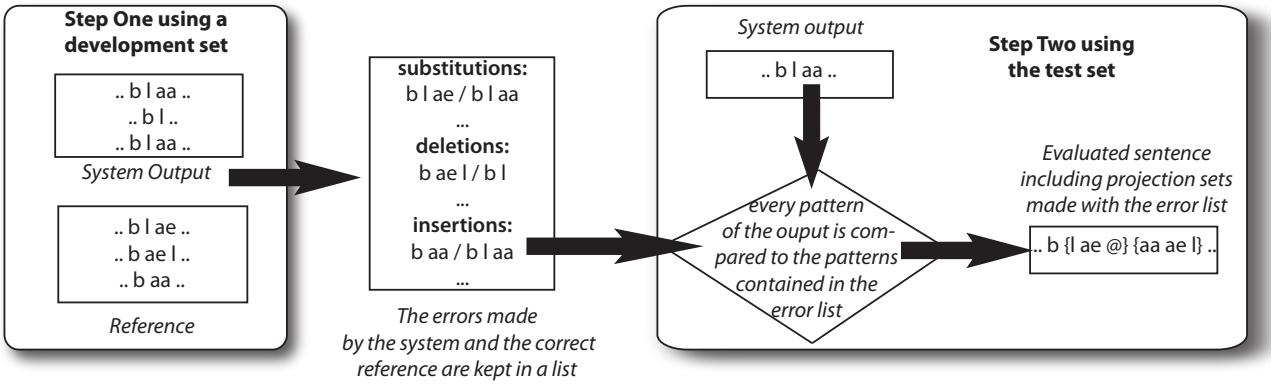

Figure 1: The output sequence of phonemes takes into account what errors have been made on a previous experiment evaluated on a development set. All those errors provide a projection set for each outputed phoneme.

follow each other. The results are summarized in table 5.

Recognition System	Sphinx 1-best	YASPER VPV	**Sphinx Context**
PER	36.2%	31%	**26.1%**

Table 5: Results at phoneme level for the two phoneme recognition systems with the context-based projection sets

4.3.5 Integrating contextual factors and PS sets

In order to tailor the projection sets based on phonetic similarity, contextual factors were integrated with the PS sets. Each phoneme is associated with a particular PS set. When a substitution is found in the development set for a particular phoneme and context, this substitution is added only if it appears in the PS extended set. When a deletion is found, the deleted phoneme is added with its PS set. If there was an insertion, the symbol @ is added to the projection set. The results are shown in table 6.

Recognition System	Sphinx 1-best	YASPER VPV	**Sphinx Merging**
PER	36.2%	31%	**19.3%**

Table 6: Results at phoneme level for the two phoneme recognition systems with the integrated projection sets with contextual factors

5. DISCUSSION AND CONCLUSION

The comparison of the performance of YASPER and the Sphinx-based phoneme recognition system using the VPV projection set provided the original motivation for the use of projection sets. Although the projection sets consist of sets of phonemes and can be regarded as classes, it is not class accuracy which is being measured; all recognition experiments were carried out with respect to individual phonemes within a class and not with respect to the classes as a whole. Two principled approaches based on phonological and phonetic similarity were presented. Each produced demonstrably better results in terms of PER than a system which only produced the most probable phoneme sequence. However, as expected, the improvement is at the expense of precision (i.e. there are many more hypotheses from which the correct hypothesis must be chosen by scoring). In order to reduce the size of the projection sets another factor needs to be included, namely contextual information. An initial experiment to identify projection sets based only on errors in specific contexts also produced promising results in terms of PER. Further experiments will be conducted to identify the most appropriate similarity sets based on inheritance hierarchies such as those suggested by [7]. The next step in the development of this approach is to use the contextual information more explicitly to tailor the projection sets based on phonetic similarity. One possible approach to the integration of PS and contextual factors has produced a phoneme error rate of 19.3%. This approach to phoneme recognition provides an alternative to the YASPER system mentioned above. However, rather than regarding this approach as a competing system, the intention is to develop a model in which the two systems will provide alternative parallel streams of hypotheses for each interval of the speech signal and will work together as experts thus facilitating robust recognition.

REFERENCES

[1] D. Aioanei. *YASPER, a knowledge-based and data-driven speech recognition framework*. PhD thesis, University College Dublin, 2008.

[2] D. Aioanei, M. Neugebauer, and J. Carson-Berndsen. Efficient phonetic interpretation of multilinear feature representations for speech recognition. In *Proceedings of the 2nd Language & Technology Conference*, pages 3–6, Poznan, Poland, 2005.

[3] J. Carson-Berndsen. *Time Map Phonology: Finite State Models and Event Logics in Speech Recognition*. Kluwer Academic Publishers, Dordrecht, Holland, 1998.

[4] J. G. Fiscus. A post-processing system to yield reduced word error rates: Recognizer output voting error reduc-

tion (ROVER). In *IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 347–352, Santa Barbara, USA, 1997.

[5] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, and N. Dahlgren. The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CDROM. 1993.

[6] A. Halberstadt and J. Glass. Heterogeneous acoustic measurements for phonetic classification. In *Proceedings of Eurospeech*, Rhodes, Greece, 1997.

[7] M. Neugebauer. *Constraint-based acoustic modelling*. Peter Lang Publishing, Frankfurt am Main, 2007.

[8] P. Scanlon, D. Ellis, and R. Reilly. Using broad phonetic group experts for improved speech recognition. *IEEE Transactions on Audio, Speech and Language Processing*, 15(3):803–812, 2007.

[9] P. Schwarz, P. Matejka, and J. Cernocky. Hierarchical structures of neural networks for phoneme recognition. In *Proceedings of ICASSP*, volume 2, pages 325–328, Toulouse, France, 2006.

[10] Speech recognition scoring toolkit (sctk), 1997. http://www.nist.gov/speech/tools/.

[11] The CMU-Cambridge Statistical Language Modeling toolkit, 1999. http://www.speech.cs.cmu.edu/SLM/toolkit.html.

[12] CMU Sphinx decoder 3.7. The Sphinx Group at Carnegie Mellon University, 2007. http://cmusphinx.sourceforge.net/.

[13] S. Young. The general use of tying in phoneme-based HMM speech recognizers. In *Proceedings of ICASSP*, volume 1, pages 569–572, San Francisco, USA, 1992.

[14] S. Zahorian, P. Silsbee, and X. Wang. Phone classification with segmental features and binary-pair partitioned neural network classifier. In *Proceedings of ICASSP*, volume 2, pages 1011–1014, Munich, Germany, 1997.