On Confusions in a Phoneme Recognizer

Andrew Lovitt [a] [b]   Joel Pinto [b] [c]
Hynek Hermansky [b] [c]

IDIAP–RR 07-10

March 2007

soumis à publication

[a] University of Illinois at Urbana-Champaign, IL., USA
[b] IDIAP Research Institute, Martigny, Switzerland
[c] École Polytechnique Fédérale de Lausanne (EPFL), Switzerland

Rapport de recherche de l'IDIAP 07-10

On Confusions in a Phoneme Recognizer

Andrew Lovitt Joel Pinto Hynek Hermansky

March 2007

soumis à publication

Résumé. In this paper, we analyze the confusions patterns at three places in the hybrid phoneme recognition system. The confusions are analyzed at the pronunciation, the posterior probability, and the phoneme recognizer levels. The confusions show significant structure that is similar at all levels. Some confusions also correlate with human psychoacoustic experiments in white masking noise. These structures imply that not all errors should be counted equally and that some phoneme distinctions are arbitrary. Understanding these confusion patterns can improve the performance of a recognizer by eliminating problematic phoneme distinctions. These principles are applied to a phoneme recognition system and the results show a marked improvement in the phone error rate. Confusion pattern analysis leads to a better way of choosing phoneme sets for recognition.

1 Introduction

The propagation of confusion patterns through speech recognition systems is a largely unanalyzed facet of hybrid hidden Markov model - artificial neural network (HHM-ANN) [1] speech recognizers. Confusions originate at all levels of the recognizer and cascade through the following stages. Pronunciation errors originate at the speaker where words with similar phonetic structure will be easily substituted. The second source of confusion patterns are from errors inherent to the recognizer itself. The confusions are evident at the posterior probabilities, which is the output of the artificial neural network, and the actual phoneme recognizer.

The confusion patterns illustrate systematic differences in the interpretation of the errors in a phoneme recognizer. Confusion pattern analysis has been used in many experiments to understand how humans confuse phonemes [2]. Some phonemes are confused at all levels of the phoneme recognizer. When the confusions are systematic they illustrate flaws in the recognizer that can be easily understood. For instance if the recognizer confuses all voiced consonants with all unvoiced consonants, this would show that the recognizer needs improvement in the detection of voicing. The confusion patterns also provide a way of understanding which confusions could be ‘ignored’ since they are likely not actual errors. These confusions may be due to many issues such as improper alignment of the phonemes. In this way not all errors are created equal.

2 Phoneme Recognizer

The phoneme recognizer is trained to recognize phonemes from the TIMIT corpus. The recognizer is built up from a Multi-resolution RASTA [3](MRASTA) feature extraction, artificial neural network, and a viterbi phoneme recognizer. All speech has a sampling rate of 8kHz. The system is seen in fig. 1.

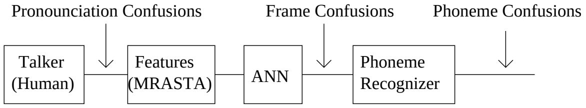

Fig. 1 – The block diagram of the system analyzed. The figure shows the points where the confusion matrices are analyzed.

2.1 Multi-resolution RASTA

The MRASTA feature extraction is used to obtain the feature vector. Features are generated every 10 ms from the acoustic signal (denoted as 1 frame). Critical band spectral analysis (Auditory Spectral Analysis step in the PLP technique [4]) is first performed on the speech signal with a window length of 25 ms and a step size of 10 msec. The resulting critical band spectrogram is then filtered using a bank of 2-D filters with varying temporal resolution to obtain a 448 dimensional feature vector for every frame.

2.2 Artificial Neural Network (ANN)

The features from the MRASTA front-end the input to a multi-layer perceptron (MLP) neural network. The MLP is implemented and trained using Quicknet [5]. This MLP is trained to discriminate between the phonemes. The MLP produces posterior probabilities for every frame where each element is the probability that the corresponding phoneme was spoken during that frame. The MLP was trained using hard labels. It contained 1 hidden layer with 800 neurons. The resultant MLP trained to 71.66% accuracy for the training data when 10% of the training data was used as cross-validation.

The labels were generated from the ‘.phn’ files provided by the corpus. In situations where there is a phoneme boundary resided between two 10 ms delineators, the boundary is rounded to the nearest 10 ms. The TIMIT corpus provided a phoneme set of 61 phones. From this large set some phonemes were merged, thus two silence label were created. One label was for regions with a long stretch of silence (‘h#’). A second label marked the phonemes where there was a short silence in the middle of words or a sentence. These phones were ‘epi’, ‘pau’, and all the closure silences. This produced a label set of 54 phones of which two were ‘silence’ classes.

2.3 Viterbi Phoneme Recognizer

The phoneme recognition decoding network consists of context independent phonemes in parallel with uniform entrance probability. Each phoneme is modeled as a left-right HMM with 3 emitting states forcing a minimum duration of 30 ms for each phoneme. The self and next state transition probability is fixed at 0.5 each. The emission likelihood for the HMM state is the phoneme posterior probability from the MLP scaled by the phoneme prior probability. In our experiments, uniform prior probability is assumed. The phoneme sequence was recognized by the Viterbi algorithm on the decoding network. This was implemented using the NOWAY software package developed at ICSI.[1]

For this analysis all silence phones are ignored since the actual phoneme sequence is of primary interest. The training data had a 37% and the test data had a 41.8% phone error rate.

2.4 Phoneme Alignment

A weighted Levenshtein algorithm [6] is used to evaluate and align the substitution, insertion, and deletion errors. This algorithm leverages knowledge of the substitution confusion patterns in the data to more accurately align the phones of the target and hypothesis. This produces confusion matrices which are more representative of the confusion patterns and eliminates the noise in the confusion matrix caused by improper alignment. This algorithm find the hypothesis in multiple hypothesis situations which has the most common substitutions.

3 Confusion Patterns

3.1 Language Level Confusions

The pronunciation confusion matrix for TIMIT is made by comparing the pronunciation of each word spoken with the official dictionary pronunciation (provided by TIMIT). There is a handful of words which are not in the TIMIT dictionary and are thus not included in the pronunciation confusion matrix. For example ‘she’ is transcribed in the TIMIT dictionary as ‘sh iy’. However there are many phonetic pronunciations in the corpus including ‘sh ix’, ‘sh ax-h’, ‘sh ih’, ‘sh q ix’, and ‘s uw’. The most common pronunciation, ‘sh iy’, (and the correct pronunciation) has approximately 90% of the pronunciations. A list of the major confusions for each phoneme is displayed in the first column in table 1.

English has various pronunciations of every word. This may be due to many factors including the energy exerted by the speaker to articulate the phones, similarities in articulatory features between phonemes, and dialect. For each phoneme there is a small subset of phones which are possible pronunciation confusions from the pronunciations in the TIMIT corpus.

The official TIMIT dictionary did not contain ‘nx’ and ‘dx’. Both phonemes are flaps and are confused with ‘n’ or ‘t’ respectively. Also ‘en’ is mispronounced as ‘ix’. The ‘en’, ‘eng’, and ‘em’ phones are actually two phones masquerading as one, thus there are confusions with the vowels that are similar to the vowel part and the consonants which are confused with the consonant part. For example with ‘en’ there is many confusions with ‘ix’ and ‘n’. Thus in the pronunciations it is highly likely that the syllabics are only partially pronounced and thus easily confused. Additionally, these confusions are influenced by the biases of the linguists who labeled the corpus. For instance the distinction between labeling a section of speech ‘en’ or labeling it ‘ix’ followed by ‘n’ may be more a transcripter bias than an actual distinction in the phoneme labels.

[1] See http ://www.icsi.berkeley.edu/ for more information.

Phone	Pronunciation Confusions	Frame Confusions	Phoneme Confusions
iy	**ix**, **ih**	**ix**, ey, **ih**	**ix**, **ih**, y
ih	**ix**, **iy**, ax, **eh**	**ix**, **eh**, **iy**	**ix**, **eh**, **iy**
eh	**ih**, **ix**	*ae*, **ih**, *ah*, **ix**	**ih**, *ae*, *ah*, **ix**
ae	**eh**, ix	**eh**, *aw*, *ay*	**eh**, ah, *ay*, *aw*
ix	**ih**, **ax**, en, **iy**	**ih**, **ax**, **iy**	**ih**, **ax**, **iy**
ux		*uw*, *ih*, *ix*, *iy*	*uw*, *ix*, *ih*, *iy*
ax	**ix**, **ah**, ih	**ix**, **ah**	**ix**, **ah**
ax-h		*ix*, t, *ax*	*ix*, *ax*, p
uw	**ux**, ix, **uh**	**ux**, **uh**, *l*	**ux**, *l*, **uh**
uh	ix, er, **ax**	**ax**, ih, ix	**ax**, ux, ah
ah	**ax**, ix	*eh*, *aa*, **ax**	**ax**, *eh*, *aa*
ao	**aa**	**aa**	**aa**
aa	**ah**, **ao**	**ao**, *ay*, **ah**	**ao**, **ah**, *ay*
er	**axr**, ax, **r**	**axr**, **r**, *eh*	**axr**, **r**, *eh*
axr	**er**, **r**, **ax**, **ix**	**er**, **r**, **ix**, **ax**	**er**, **r**, **ix**, **ax**
ey	**eh**	*iy*, *ih*, **eh**	*ih*, *iy*, **eh**
ay	**aa**	**aa**, ah, ae	**aa**
oy	**ao**, **ow**	**ao**, **ow**	**ao**, r, **ow**
aw	**aa**	*ae*, **aa**, *ow*, ay, *eh*	*ae*, **aa**, *ow*, *eh*, l
ow	ax, uh	*l*, *ao*, *ah*	*l*, *ah*, *ao*
p		*t*, *k*, *f*	b, *t*, *k*, *f*
t	dx, q, d	*k*, *p*, ch, s	d, *p*, *k*
k		*t*, *p*, *g*	*t*, *g*, *p*
q			
b	**v**	*p*, **v**, dh	*p*, d, **v**
d	dx, **t**	**t**, jh, *g*, *dh*	*g*, **t**, *dh*
g		*k*, *d*, t	*d*, *k*
m	**em**	*n*, **em**	*n*, **em**
n	nx, en	*m*, *ng*, en	*m*, *ng*, nx
ng	**n**	**n**, *m*	**n**, *m*
nx		*n*, *dx*, *m*	*n*, *dx*, *m*
dx		*dh*, *v*, *n*, *nx*	*dh*, *nx*, d, *v*, *n*
f		*s*, *th*, *v*, z	*s*, *th*, *v*
th	**dh**, t	*f*, v, **dh**	*f*, t, **dh**, b
s	**sh**, **z**	**z**, *f*, **sh**	**z**, *f*, **sh**
sh		*s*, *zh*, *ch*	*s*, *zh*, *ch*
v	f	*dh*, *z*, f	*dh*, *z*, b
dh	th, d	v, f, th	d, b
z	**s**, zh	**s**, *v*	**s**, *v*, zh
zh	jh, z, **sh**, ch	**sh**, z, s	ux, **sh**, en
ch	**sh**	**sh**, *jh*, *t*, *s*	*t*, *jh*, **sh**, *s*
jh	zh	z, *ch*, zh	y, d, *ch*, t
l	**el**	*ow*, *w*, **el**	*w*, **el**, *ow*
r	**axr**, **er**	**er**, **axr**	**axr**, **er**
y	ix, ux	*iy*, *ih*, ux	*iy*, *ih*
w		*l*, *ao*, *uw*	*l*, *ao*, *uw*
em	m	m, uw, ax, *n*, en	ux, *n*, w
en	**ix**, **n**	**n**, **ix**, *m*, ng	**ix**, **n**, *m*
eng			
el	**l**	*ow*, **l**, *ao*, *ax*	**l**, *ow*, *ax*, *ao*
hv		*hh*	*hh*
hh	**hv**	**hv**, *q*, f	*q*, **hv**

Tab. 1 – Major confusions from all stages of the phoneme recognizer. Only the major confusions for each phoneme are shown. The phonemes are in order of probability for their respective columns. Many low probability confusions were eliminated (for space reasons) however the majority of the total number of confusions are represented for each phone. The italic blue phonemes are phones which are confused at all stages analyzed of the phoneme recognizer. The bold red phonemes are major confusions which appeared only in the posterior probability and phoneme recognizer confusions.

The insertions and deletions in the pronunciation dictionary contain systematic patterns as well. The pronunciation dictionary has very few insertions and they are overwhelmingly ‘q’ insertions. The phoneme ‘q’ is defined in the TIMIT literature as :

> glottal stop q, which may be a allophone of t, or may mark an initial vowel or a vowel boundary.

This description gives the impression that in transcribed speech there may be a lot of ‘q’ insertions whereas in the official TIMIT dictionary we expect few or none. The most deleted phones are ‘uh’, ‘p’, ‘t’, ‘k’, ‘b’, ‘d’, ‘g’, ‘n’, ‘m’, and ‘r’. The lack of a significant number of insertions also shows that while the listeners make insertions and deletions while speaking, these errors are predominately deletions (over 5 :1).

These confusion patterns show that not all pronunciation errors are created equal. A deletion of a ‘b’ consonant should not be given the same weight as deletion of a ‘s’ consonant because ‘s’ is hardly ever dropped in pronunciations whereas ‘b’ is deleted a significant amount. Thus if a speech recognizer drops a ‘t’ or substitutes an ‘ix’ for an ‘en’ the error may not be an actual error as opposed to when ‘s’ is recognized as an ‘n’.

3.2 Frame confusions

There is a difficulty in constructing confusion matrices for posterior probabilities because most frames near the boundaries between phonemes show considerable overlap with the preceding or proceeding phonemes. This means that in the top 5 posteriors (ranked by probability) for each frame near the boundary, both phones are usually present for 10-20 ms in both directions of the boundary. This effect causes confusions with the previous or following phonemes. These confusions are not actual confusions, they are just the improper labeling of the frames. It is likely that the MLP is correctly identifying the frame but the label for the frame is incorrect. Due to this, the confusion matrix for frames is made after throwing out all frames within 10 ms of the boundaries. The frame confusions are shown in the second column of table 1.

The frame confusion matrix contains more confusions than the pronunciation confusion matrix. Additionally, there are slightly different confusions however there is also large similarities as well. The largest differences between the pronunciation and frame confusion patterns are in the consonants and semi-vowels. The consonants show more confusions in the frame confusion matrix than in the pronunciation confusion matrix. The confusion patterns between the vowels are much more similar.

3.3 Phoneme Recognizer Confusions

The confusion matrix for the phoneme recognizer is made from the weighted Levenshtein aligned strings [6]. The phoneme recognition is performed on each sentence individually. The confusions for the phoneme recognizer are shown in column 3 in table 1. The insertion and deletion probabilities are very similar to the probabilities from the pronunciation confusions. There are a significant amount of ‘axh’ and ‘b’ insertions which are not seen in the pronunciation insertion patterns. However in isolation of these differences the insertions and deletions where not vastly different from the pronunciation confusion results.

4 Comparison of Confusion Patterns

4.1 Frame and Phoneme Confusions

The confusion patterns from the phoneme recognition are not significantly different from the confusion patterns seen at the frame level. All of the confusion patterns which are similar between frame confusion patterns and phoneme recognizer confusions are highlighted as *red italics* and **blue bold** in table 1. The errors pass through the phoneme recognizer for the most part without significant additional errors. Whereas the similarities between the pronunciation and frame confusions are primarily just the vowels, the confusions between the frame and phoneme recognizer confusions also include the majority of the consonants.

4.2 Common Confusion Patterns

The confusions from the frames and the phoneme recognizer show strong similarity with the confusion patterns seen in the pronunciation confusion matrix for some phonemes. These show up as **bold blue** phonemes in table 1. The similar confusions show that the phoneme recognizer confuses phones which are likely to be mispronounced. The pronunciation confusions are mostly between phones which have similar articulatory structure. Thus the MLP was unable to completely distinguish the phones that are produced very similarly. Additionally, it is very likely that the phonemes in the corpus are not the strongest examples of the phones they labeled as, thus the MLP is likely trained on cases where a phone that could easily be either of the confused phones. This lead to a lack of distinction between similar phonemes by the phoneme recognizer.

Some similarities show places where the articulatory differences in the phones are likely to be very small and arbitrary. For instance 'er', 'axr', and 'r' are all highly confused with each other. These phonemes are likely so similar that the distinction between them could be eliminated. However in modern speech recognizer analysis the misplacing of a 'r' with 'axr' would be counted as an error. The more correct analysis is that all three phonemes are too similar to distinguish thus they are not errors. Another case where this is evident is distinction between 's' and 'z'. In the phonetic transcriptions the distinction between 's' and 'z' is arbitrary especially when the consonant ends a word. In these cases it is not important whether 's' or 'z' is said but whether one of them was said.

For example if the target string is 'sh iy hv ae d' and the ANN posteriors reported that the string was 'zh ix hh ae d', it is not a completely impossible situation that the reported frame posteriors are actually correct because 'sh' and 'zh', 'iy' and 'ix', and 'hv' and 'hh' are pronunciation and phoneme recognizer confusions. Thus in this case the recognizer should not count these as errors or count them are trivial errors. The differences between these phonemes are probably not trainable since they may not exist. However, if the phoneme recognizer reported 't ae q ae d' the errors are not similar to the confusions and thus show actual recognition errors.

4.3 Comparison with Human Confusions

There are similarities between the confusion patterns seen in humans as white noise is raised [7], [2] (MNR) and the confusions in table 1. The nasal and the 'p', 't', and 'k' confusions are similar between the repeat of Miller-Nicely and the phoneme recognizer. There are also similarities in the amount of voicing confusions. In the MNR experiment there are significant voicing confusions for fricatives. These error patterns are seen in the confusions for 's', 'z', 'zh', and 'sh'. The second type of error found in both sets of results are errors in the articulatory feature of place. The place errors however are very similar between the two experiments. The confusions from the phoneme recognizer show that the articulatory features most likely to be lost are voicing and place. The confusion patterns of the phoneme recognizer are most similar to the confusion patterns at -6 dB SNR from the MNR experiment [7].

These results imply a similarity in the events extracted by humans and by machine recognition. The difference is that the event extraction is less robust for the phoneme recognizer. The events that are less robust to noise are the events the phoneme recognizer is not properly recognizing and is confusing. Additionally, the distinctions that are most robust in noise for humans are the distinctions where the phoneme recognizer does very well. This implies that research should be focused to investigate the events which humans do not recognizer at high SNRs.

Phone Groups	Phone Groups
iy, ix, ih	hv, hh
ax, ah	ng, n, nx, en
ae, eh	th, dh
uw, ux	s, z
ao, aa	em, m
er, axr, r	b, v
zh, sh	el, l
short and long silences	

Tab. 2 – This table shows 15 sets of phonemes which have prolific confusion patterns so the distinction is assumed to be arbitrary. Thus errors between members in a group are not counted as errors in the recognition class evaluation.

	Pronunciations (% correct)	Frame P_c w/o silence	Phoneme Recognizer (PER)
52 Phone	34.02%	60.4%	41.8
Collapsed Phone	49.40%	68%	34.2

Tab. 3 – This table shows the results of the frame P_c and the phoneme recognizer PER on the test data. Additionally the number of pronunciations which are correct given the recognition groups is reported. The data is shown for both the original 52 phone set (+ two silence phonemes) and the collapsed phone recognition classes (35 groups).

5 Application

To illustrate the fact that not all errors are the same, the phoneme recognizer is reanalyzed with a smaller subset of phonemes. The most common confusions of phonemes from table 1 across all confusions matrices are collected into recognition classes. These sets of phones are derived from table 1, and are seen in table 2. All phonemes not included in table 2 are given their own recognition class. This is done because the confusions between these phones are prolific at all levels of the phoneme recognizer. This collapsed set had 35 recognition classes of which one is ‘silence’.

For the collapsed phones the posteriors of the original MLP net are added together based on the recognition groups stated in the table. After the posteriors are collapsed the phoneme recognition is performed. An important metric for collapsing the phones into recognition classes is the amount of collisions the recognition classes cause in the dictionary. The recognition classes only add collisions to 1.67% of the words in the TIMIT dictionary. Thus there is only about 47 pairs of words which are impossible to distinguish given the new recognition groups. The results of recognition given the recognition classes is seen in table 3. This shows that given sensible recognition classes based on the confusion patterns the mispronunciations can be reduced as can the overall errors in the phoneme recognizer without significant degration of the ability to identify the words from the dictionary.

Additionally applications of the principles of confusion matrix analysis can be seen in [8]. In this work the confusion patterns are leveraged to improve the keyword spotting ability.

6 Conclusions

The confusion patterns in both the pronunciation and the output of a phoneme recognizer are very similar. The phoneme recognizer is making errors similar to the errors a human speaker make in production. Due to this, there is significant structure which can be exploited in the confusion patterns to design phoneme classes. These classes can help eliminate errors and only trivially increasing dictionary confusions. The confusion patterns for certain phonemes exhibit confusion patterns which are reminiscent of the confusions patterns for human recognition in masking noise. Thus the features human lose in white masking noise is similar to those lost in the phoneme recognition. Analysis shows that some errors are not errors and should not be counted as such. Some distinctions in phonemes add to the PER or P_e needlessly. Thus, the interpretation of results should take into account whether the types of errors based on the confusion patterns.

7 Acknowledgments

This work was done under the DIRAC integrated project funded by the European Community.

Références

[1] Herve Bourlard and Nelson Morgan, *Connectionist Speech Recognition - A Hybrid Approach*, Kluwer Academic Publishers, Boston, 1994.

[2] George A. Miller and Patricia E. Nicely, "An analysis of perceptual confusion amoung English consonants," *Journal of the Acoustical Society of America*, vol. 27, pp. 338–352, 1955.

[3] Hynek Hermansky and Petr Fousek, "Multi-resolution rasta filtering for tandem based asr," in *Proceedings of International Conference on Spoken Language Processing (ICSLP)*, September 2005, pp. 361–364.

[4] Hynek Hermansky, "Perceptual liner predictive (PLP) analysis of speech," *Journal of the Acoustical Society of America*, vol. 87, pp. 1738–1752, April 1990.

[5] David Johnson et al., "ICSI quicknet software package.," 2004, http ://www.icsi.berkely.edu/Speech/qn.html.

[6] Andrew Lovitt, "Correcting confusion matrices for phone recognizers," IDIAP-COM 03, IDIAP, 2007.

[7] Andrew Lovitt and Jont Allen, "50 years late : Repeating Miller-Nicely 1955," in *Proceedings of International Conference on Spoken Language Processing (ICSLP)*, September 2006.

[8] Joel Pinto, Andrew Lovitt, and Hynek Hermansky, "Exploiting phoneme similarities in hybrid HMM-ANN keyword spotting.," IDIAP-RR 11, IDIAP, 2007.