Investigating phonetic information reduction and lexical confusability

William Hartmann[1], *Eric Fosler-Lussier*[1,2]

Department of Computer Science and Engineering[1]
Department of Linguistics[2]
The Ohio State University, Columbus, OH, USA

hartmann.59@osu.edu, fosler@cse.ohio-state.edu

Abstract

In the presence of pronunciation variation and the masking effects of additive noise, we investigate the role of phonetic information reduction and lexical confusability on ASR performance. Contrary to previous work [1], we show that place of articulation as a representation for unstressed segments performs at least as well as manner of articulation in the presence of additive noise. Methods of phonetic reduction introduce lexical confusibility which negatively impact performance. By limiting this confusability, recognizers that employ high levels of phonetic reduction (40.1%) can perform as well a baseline system in the presence of nonstationary noise.

Index Terms: spoken language analysis, speech recognition, articulatory features

1. Introduction

While pronunciation dictionaries usually have one or more phonetically based, canonical pronunciations for each entry, the actual pronunciations of a speaker in both the spontaneous and read speech conditions can vary greatly. A common approach is to model this as pronunciation variation [2] where the pronunciation is predicted from the data or explained by a set of rules. Perceptual studies such as Miller and Nicely [3], which examine phonetic confusions, are often used as a basis of argument for these methods.

In linguistics these variations are typically handled as rules regarding articulatory features rather than a complete change in the phone. For example, /l/ can become devoiced when it follows an unvoiced stop like /p/ as in the word *please*. In that case, the place and manner of articulation stay constant while the voicing changes. The call for the use of articulatory features is certainly not new [4, 5]; recent research has shown that the incorporation of these features can improve the performance of automatic speech recognition [6, 7].

Instead of incorporating articulatory features as additional features in the system, we can apply them more directly to the problem of pronunciation variation. Briscoe [1] provides an insight into how this can be accomplished through his work in lexical access in connected speech. He showed that the strategy of replacing the phones in unstressed syllables with their manner of articulation was most effective at producing a minimal list of word candidates that still contained the target word in the presence of pronunciation variation. Manner of articulation was used because he claimed detection of place of articulation would be more affected by the presence of noise.

We partially investigated his claims in [8] where we replaced the phonological representation in portions of the unstressed syllables with manner of articulation. This effectively removed the problem of handling pronunciation variation by representing words only with the phonological features that were least expected to vary. The initial expectation was that any recognizer which used a reduction of phonetic information would perform worse than a baseline system using a full phonetic representation. Surprisingly, we found that we could reduce a significant number of the phones and still obtain performance that did not differ significantly from the baseline and could actually outperform the baseline in the presence of stationary noise.

As in the previous study, our goal is not necessarily to produce a better speech recognizer, but rather to determine how much linguistic information we need to obtain from the speech signal in order to produce comparable results to a standard system. In this work, we extend the results of the previous study in three ways and further investigate the amount of linguistic information conveyed in unstressed syllables through the use of automatic speech recognizers. While [8] shows that manner of articulation is robust in the face of car noise, it does not test Briscoe's statement that this representation is less masked by noise than place of articulation. We examine the use of place of articulation as a representation for unstressed pronunciation reduction and compare the results to those of manner of articulation. In keeping with the idea of improving lexical access, we also investigate the unavoidable consequence of increased lexical confusability from the reduction of phonological information. Finally, we further explore the effects of noise by examining various signal to noise ratios (SNRs).

In Section 2 we present the methodology used for the training and testing of the recognizers. The results are discussed in Section 3 and our conclusions are presented in Section 4.

2. Experimental Setup

2.1. Baseline

Using the HTK [9] system, we trained a baseline recognizer on the speaker-independent portion of the Wall Street Journal 5k task (WSJ0). The CMU dictionary was used for our baseline pronunciations; tied-state inter-word triphones with 16 Gaussians per state comprised the acoustic model using 39 phones. PLP features were mean subtracted and delta and double-delta coefficients were used, giving a 39-dimensional feature vector.

In addition to clean speech, three separate noise cases (car, babble, and factory) were tested where noise from the NOISEX database [10] was added to the clean speech at various SNRs.[1]

[1] While additive noise does not account for the Lombard effect [11], where speakers do emphasize speech to overcome noise, there are situ-

Baseline	s	ah	l	uw	sh	ah	n
Manner	fricative	vowel	approximant	vowel	fricative	vowel	nasal
Manner_unstressed	fricative	vowel	l	uw	fricative	vowel	nasal
Manrime_unstressed	s	vowel	l	uw	sh	vowel	nasal
Mancoda_unstressed	s	ah	l	uw	sh	ah	nasal
Place	dental	mid vowel	lateral	back vowel	alveolar	mid vowel	alveolar
Place_unstressed	dental	mid vowel	l	uw	alveolar	mid vowel	alveolar
Placerime_unstressed	s	mid vowel	l	uw	sh	mid vowel	alveolar
Placecoda_unstressed	s	ah	l	uw	sh	ah	alveolar
Place_1_vowel	dental	vowel	lateral	vowel	alveolar	vowel	alveolar
Manner_3_vowels	fricative	mid vowel	approximant	back vowel	fricative	mid vowel	nasal

Table 1: Example pronunciations for the word *solution* for the articulation based recognizers

The car noise is the recorded interior noise of a Volvo 340 traveling at a constant rate of speed on the highway and is considered stationary. Babble noise was recorded in a canteen containing 100 people. Factory noise was recorded in a factory near plate-cutting and welding equipment. Both babble and factory are considered nonstationary noise.

2.2. Articulation Based Recognizers

Ten alternative dictionaries were derived from the CMU dictionary; these modified dictionaries were used for the training and testing of the articulation based recognizers in the same manner as the baseline recognizer. The phonetic representations were mapped to their corresponding articulatory features to produce the alternate dictionaries. Each dictionary used either place or manner of articulation. Four methods were used to determine which portions of the phonetic representation to replace with articulatory features based on lexical stress and syllable position.

The simplest case was a reduction of all phones to their articulatory class as in the *Place*, *Manner*, *Manner_3_vowels*, and *Place_1_vowel* recognizers, where *Manner_vow* uses three classes for vowels and *Place_1_vowel* uses only one class for vowels. As discussed in Section 3, they were included to allow for a more fair comparison between manner and place of articulation since place has three vowel classes and manner only has one.

The other three methods replaced a portion of the unstressed syllable. We expected the unstressed syllable to contain more variation [12, 13] and stressed syllables are more likely to contribute to lexical access [1]. Although the CMU dictionary provides stress markings, it does not provide the syllables for the pronunciations. We used the NIST tsylb2 syllabification program [14] for determining the syllables in the pronunciations. Both the *Manner_unstressed* and *Place_unstressed* recognizers replaced the entire unstressed syllable with its respective articulation feature. The rime (non-onset) of the unstressed syllable was replaced in *Manrime_unstressed* and *Placerime_unstressed*. Only the coda was replaced in *Mancoda_unstressed* and *Placecoda_unstressed*. An example of the respective pronunciations for a given word are shown in Table 1.

2.3. Minimum Confusibility Recognizers

Since there are fewer articulation classes than the phone classes, this reduction procedure introduced further confusability into the dictionaries. We can rank the recognizers in Table 2 by their performance on clean speech. The best performing recognizers also happen to contain the fewest number of confusable words in the dictionary. In order to further examine the role of confusability, the dictionaries of the *minimum confusability recognizers* were created such that the pronunciation of each word was only reduced if it didn't cause the pronunciation to be identical to another word in the dictionary [15].

Reduced phones are defined as phones that have been replaced by their articulation class in the dictionary. The minimum confusability recognizers have fewer phones reduced compared to their standard counterparts because the rules for reduction are no longer universally applied to every word. For example, in the case of place of articulation this caused the percentage of reduced phones to go from 100% to 70.6%.

3. Results

The results for the standard articulatory based recognizers and the minimum confusability versions are shown in Tables 2 and 3 respectively. The phones reduced column displays the percentage of phones in the dictionary that were replaced by either place or manner of articulation. Results of the best performing recognizers on car and factory noise at various SNR can be seen in Tables 4 and 5 where nearly every result is insignificantly different from the baseline.

As expected, the baseline recognizer performs the best in the clean speech task. Since we are dealing with unmasked, read speech, backing off to articulatory features may not be necessary or beneficial in this case. However, it is worth noting that several of the recognizers performed well enough that the difference was not statistically significant compared to the baseline (e.g. the *Placerime_unstressed.minconfuse* recognizer, with 28.4% of its phones reduced to the place of articulation, was within 0.3% of the baseline). In fact, all six recognizers which reduced less than 30% of the phones to articulatory features performed just as well as the baseline on clean speech.

Under the various noise conditions, the usefulness of the the articulatory based representations become more apparent. The results on the car case showed similar patterns to clean speech, but the nonstationary noise cases (babble and factory) showed a marked difference in performance. There are several cases where the baseline is outperformed. In addition, the *Placerime_unstressed* recognizer does not significantly differ from the

ations where speakers are unaware of the noise (e.g. car radio setting) and listeners are still able to perceive the speech.

Recognizer	clean	car 10dB	babble 10dB	factory 10dB	Phones reduced
Baseline	91.2%	89.2%	70.9%	74.3%	0%
Manner	53.3%	49.4%	30.7%	31.0%	100%
Manner_unstressed	84.7%	81.7%	61.2%	65.3%	47.9%
Manrime_unstressed	88.7%	85.2%	66.3%	70.0%	30.0%
Mancoda_unstressed	90.8%	88.8%	70.7%	73.5%	11.1%
Place	73.7%	70.1%	50.4%	54.4%	100%
Place_unstressed	88.4%	85.7%	66.8%	71.9%	47.9%
Placerime_unstressed	89.4%	87.5%	70.2%	74.2%	30.0%
Placecoda_unstressed	90.4%	88.2%	70.5%	74.4%	11.1%
Manner_3_vowels	70.5%	66.9%	45.0%	46.2%	100%
Place_1_vowel	58.9%	56.3%	39.8%	40.7%	100%

Table 2: Word accuracy results on the WSJ0 evaluation set.

Recognizer	clean	car 10dB	babble 10dB	factory 10dB	Phones reduced
Manner.minconfuse	86.5%	83.2%	62.0%	65.3%	43.8%
Manner_unstressed.minconfuse	89.8%	87.5%	69.2%	71.8%	44.2%
Manrime_unstressed.minconfuse	90.7%	88.3%	70.0%	72.9%	28.3%
Mancoda_unstressed.minconfuse	90.9%	88.4%	71.5%	74.5%	10.6%
Place.minconfuse	87.6%	84.2%	62.8%	67.2%	70.6%
Place_unstressed.minconfuse	89.9%	88.0%	70.2%	73.6%	40.1%
Placerime_unstressed.minconfuse	90.9%	88.4%	69.8%	74.3%	28.4%
Placecoda_unstressed.minconfuse	91.0%	88.6%	70.6%	75.1%	10.6%

Table 3: Word accuracy results on the WSJ0 evaluation set using the minimum confusability dictionaries.

baseline unlike the clean speech and car noise cases. Even more impressive is the result of the *Place_unstressed.minconfuse* recognizer. With 40.1% of its phones reduced to place of articulation, its performance is still near the performance of the baseline and exceeds it in several cases.

We expected that the increased lexical confusability introduced by the phonetic reduction would negatively impact results, which was borne out by the differences in the results between the standard articulatory feature and the minimum confusability recognizers. Both of the standard rime recognizers saw a significant improvement in performance with just a slight reduction in the number of phones that were reduced. The strongest example of this improvement is revealed when comparing the results of the *Place_unstressed.minconfuse* and *Placecoda_unstressed* recognizers. As we can see from Tables 2 and 3, their word accuracies do not differ significantly even though *Place_unstressed.minconfuse* has nearly four times the amount of phones reduced to place of articulation in the dictionary.

Obviously the number of reduced phones does not tell the entire story since each word in the dictionary is not equally weighted in the evaluation metric. The 10 most frequent words in the evaluation set account for 20% of the tokens in the evaluation set. If the minimum confusability based recognizers simply didn't contain articulation classes in their pronunciations, this could go a long way in explaining the improvement. However this is not the case. Using the *Placerime_unstressed.minconfuse* recognizer as an example, half of the 10 most frequent words (including the two most frequent) still use place of articulation in their pronunciation. The significant improvement that *Placerime_unstressed.minconfuse* sees over *Placerime_unstressed* cannot simply be accounted for by looking at the loss of the articulation classes in the pronunciations of the most frequent words.

Due to the conclusions of studies like [1] and our previous results in [8], we expected the recognizers based on manner of articulation to outperform those using place of articulation in the presence of noise. However, the results show that the *Place* recognizer outperforms the *Manner* recognizer by a large margin. Since we use three classes for vowels with place of articulation and only one for manner, this may not be a fair comparison. To rectify this issue, we created the *Manner_3_vowels* and *Place_1_vowel* recognizers. The *Manner_vow* recognizer used three classes for vowels just like the *Place* recognizer, but still falls short in terms of accuracy. Similarly the *Place_1_vowel* recognizer only used a single class for vowels and still outperformed the *Manner* recognizer. At lower levels of phone reduction in the clean speech and car noise cases, both representations perform equally well.

However, the recognizers using place of articulation were the best for the factory noise case leading to the conclusion that place of articulation may be less affected by nonstationary noise than manner. We again see that the *Place_unstressed.minconfuse* recognizer, which has 40.1% of its phones replaced by place of articulation in the dictionary, performs remarkably well. It is even able to match the performance of *Mancoda_unstressed.minconfuse*, the best performing manner based recognizer. Finally, we can see that the *Placecoda_unstressed.minconfuse* recognizer consistently outperforms the baseline; significantly so for the 8dB case.

4. Conclusion

Through our experiments we have shown that even when significantly reducing the phonetic information in unstressed syllables, we can obtain results comparable to the baseline. Contrary to Briscoe's [1] conclusions, we have shown that using place of articulation is just as good, if not better, than manner of articulation at handling the masking effects of noise. Even when accounting for the discrepancy in the number of

Recognizer	10dB	8dB	6dB	4dB	2dB	0dB
Baseline	89.2%	88.6%	87.8%	86.8%	86.2%	84.3%
Place_unstressed.minconfuse	88.0%	87.1%	86.1%	85.2%	84.0%	82.8%
Placecoda_unstressed.minconfuse	88.6%	88.0%	87.2%	86.5%	85.4%	84.0%
Placecoda_unstressed	88.2%	88.0%	87.3%	86.4%	85.3%	84.4%
Mancoda_unstressed.minconfuse	88.4%	87.9%	87.5%	87.3%	85.5%	84.5%
Mancoda_unstressed	88.8%	88.3%	87.4%	87.2%	85.6%	84.8%

Table 4: Word accuracy results on the WSJ0 evaluation set with car noise added.

Recognizer	10dB	8dB	6dB	4dB	2dB	0dB
Baseline	74.3%	64.5%	55.0%	40.5%	29.3%	16.2%
Place_unstressed.minconfuse	73.6%	65.6%	55.5%	41.4%	30.1%	18.5%
Placecoda_unstressed.minconfuse	75.1%	66.2%	56.1%	41.5%	29.9%	16.9%
Placecoda_unstressed	74.4%	64.6%	54.6%	40.4%	29.1%	15.7%
Mancoda_unstressed.minconfuse	74.5%	65.6%	55.2%	40.9%	28.6%	15.4%
Mancoda_unstressed	73.5%	64.9%	54.2%	38.8%	28.3%	14.7%

Table 5: Word accuracy results on the WSJ0 evaluation set with factory noise added.

vowel classes, the accuracy of the place of articulation based recognizers significantly outperformed their corresponding manner based recognizers at high levels of phonetic reduction. This refutes the idea that manner of articulation is the more noise-resistant feature relative to place. Rather, a more complex set of cues is making it through the noise. The results of the two types of articulatory feature based recognizers were more comparable at lower levels of reduction on both the clean speech and car noise tasks. Under the factory noise condition, however; the place of articulation based recognizers performed better. In fact, they produced comparable results with nearly four times the amount of phonetic reduction.

Increased lexical confusability introduced through phonetic reduction is certainly an issue. The *Place_unstressed.minconfuse* recognizer was able to perform just as well as the *Placecoda_unstressed* recognizer even with a far greater amount of phones replaced with place of articulation. Since the increased confusability makes it impossible for the system to accurately distinguish certain words, the benefits of reducing the phonetic information may be overshadowed.

In future work we will investigate Briscoe's claim that voicing is robust in the face of noise and that the results hold in both the Lombard and spontaneous speech conditions. Also, armed with a better understanding of the phonetic information conveyed by unstressed syllables, we will attempt to apply this knowledge to the problem of pronunciation variation.

5. Acknowledgment

We gratefully acknowledge William Hartmann's support under NSF grant IIS-0835396 during the course of this project; this work was also partially supported by NSF CAREER grant IIS-0643901. The opinions and conclusions expressed in this work are those of the authors and not of any funding agency.

6. References

[1] E. J. Briscoe, “Lexical access in connected speech recognition.” Proceedings of the 27th Annual Meeting of the Association for Computational Linguists, 1989, pp. 84–90.

[2] H. Strik and C. Cucchiarini, “Modeling pronunciation variation for ASR: A survery of the literature,” *Speech Communication*, vol. 29, pp. 225–246, 1999.

[3] G. Miller and P. Nicely, “An analysis of perceptual confusions among some English consonants,” *Journal of the Acoustical Society of America*, vol. 27, pp. 338–352, 1955.

[4] L. Deng, G. Ramsay, and D. Sun, “Production models as a structural basis for automatic speech recognition,” *Speech Communication*, vol. 33, pp. 93–111, 1997.

[5] M. Ostendorf, “Moving beyond the ’beads-on-a-string’ model of speech.” IEEE ASRU Workshop, 1999.

[6] K. Kirchhoff, G. A. Fink, and G. Sagerer, “Combining Acousitc and Articulatory Information for Robust Speech Recognition,” *Speech Communications*, vol. 37, pp. 303–319, 2002.

[7] S. King, J. Frankel, K. Livescu, E. McDermott, K. Richmond, and M. Wester, “Speech production knowledge in automatic speech recognition,” *Journal of Acoustic Society of America*, vol. 121, pp. 723–742, 2007.

[8] E. Fosler-Lussier, C. A. Rytting, and S. Srinivasan, “Phonetic ignorance is bliss: Investigating the effects of phonetic information reduction on ASR performance.” Proceedings of Interspeech, 2005.

[9] S. Young, G. Evermann, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book*. Cambridge University Publishing Department, 2002. [Online]. Available: http://htk.eng.cam.ac.uk

[10] A. P. Varga, H. J. M. Steeneken, M. Tomlinson, and D. Jones, “The NOISEX-92 Study on the Effect of Additive Noise on Automatic Speech Recognition,” Speech Research Unit, Defense Research Agency, Malvern, UK, Tech. Rep., 1992.

[11] J.-C. Junqua, “The Lombard reflex and its role on human listeners and automatic speech recognizers,” *Journal of the Acoustical Society of America*, pp. 510–524, January 1993.

[12] D. R. van Bergem, “Acoustic vowel reduction as a function of sentence accent, word stress, and word class,” *Speech Communication*, vol. 12, pp. 1–23, 1993.

[13] R. J. J. H. van Son and L. C. W. Pols, “An acoustic profile of consonant reduction.” Proceedings of ICSLP, 1996, pp. 1529–1532.

[14] W. Fisher, *The tsylb2 Program: Algorithm Description*, NIST, part of the tsylb2-1.1 software package, 1996.

[15] T. Sloboda and A. Waibel, “Dictionary learning for spontaneous speech recognition,” vol. 4. ICSLP, 1996, pp. 2328–2331.