TRANSMISSIONS AND TRANSITIONS: A STUDY OF TWO COMMON ASSUMPTIONS IN MULTI-BAND ASR

Nikki Mirghafori and Nelson Morgan

International Computer Science Institute, 1947 Center St, Berkeley, CA 94704
University of California at Berkeley, EECS Department, Berkeley, CA 94720
Tel: (510) 643-9153, FAX: (510) 643-7684, Email: {nikki, morgan}@icsi.berkeley.edu

ABSTRACT

Is multi-band ASR inherently inferior to a full-band approach because phonetic information is lost due to the division of the frequency space into sub-bands? Do the phonetic transitions in sub-bands occur at different times? The first statement is a common objection of the critics of multi-band ASR, and the second, a common assumption by multi-band researchers. This paper is dedicated to finding answers to both these questions.

To study the first point, we calculate phonetic feature *transmission* for sub-bands. Not only do we fail to substantiate the above objection, but we observe the contrary. We confirm the second hypothesis by analyzing the phonetic *transition* lags in each sub-band. These results reinforce our view that multi-band speech analysis provides useful information for ASR, particularly when band merging takes place at the end state for a phonetic or syllabic model, allowing sub-bands to be independently time-aligned within the model.

1. INTRODUCTION

There has been much interest generated in the speech recognition community on multi-band ASR since Jont Allen's cogent retelling of Harvey Fletcher's work on articulation index [4, 1]. The main idea of this approach is to divide the signal into separate spectral bands, process each independently (typically generating state probabilities or likelihoods for each), and then merge the information streams, as shown in Figure 1. Some of the motivations for this multi-band approach are:

- If the speech signal has different signal-to-noise ratios per band, multi-band ASR shows graceful degradation [2, 10].
- It has been posited that acoustic evidence for sound unit identities occur at different times in different parts of the spectrum, particularly in the presence of reverberation or unusually slow or rapid speech.
- Statistical modeling may be improved by simpler and less variable signals and lower dimensionality of the feature set.
- Rao and Pearlman [9] have proven theoretically, and shown with simulations, that auto-regressive spectral estimation from sub-bands offers a gain over full-band auto-regressive spectral estimation.
- Multi-band ASR is well suited for taking advantage of parallel architectures.
- Human speech perception may be similar [4, 1].

The most common objection to the use of separate statistical models for each band has been that important information in the form of correlation between bands may be lost. Our experience and that of our colleagues has been that recognition performance has not been hurt by this approach, but in the work reported here we examine the estimator performance in a more detailed fashion. In particular, we analyze the phonetic feature transmission pattern in each sub-band, the merged multi-band, and full-band probability streams. As discussed in Section 3, we use methods similar to those of Miller & Nicely [7] and calculate confusion matrices for phone and feature classes, and use mutual information as a measure of information transmission in a channel.

In the second part of the paper, we focus our attention on the following: some multi-band researchers [2, 10, 8] have postulated that transitions in sub-bands occur asynchronously, and that a phone or syllable level merging of multi-band streams is necessary to permit independent alignment for each band within the merged unit. However, this hypothesis has not been analyzed; neither has there been a study of transition boundary shifts in the presence of speech signal variations (such as room reverberation or speaking rate). Without such evidence, we could not justify consideration of longer-term merging units for multi-band ASR. In Section 4, we examine this assumption by analyzing the transition lags in each sub-band to see if sub-band transitions occur asynchronously.

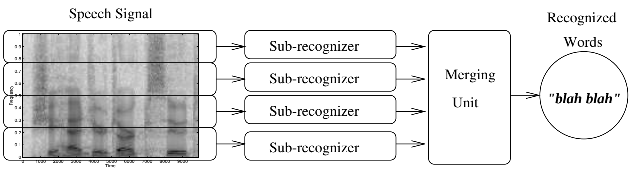

Figure 1: A simple overview of multi-band.

2. DATABASE & SYSTEM DESCRIPTION

We use the Oregon Graduate Institute NUMBERS95 database, which comprises continuous digits and numbers recorded over the telephone as a part of census data collection. The database is phonetically hand-transcribed. For the purposes of this study, we use approximately two hours of the database for training and cross validation, and forty minutes as a test set.

Our baseline full-band system is an HMM/MLP based [3] system. We train the MLP phonetic probability estimator on a nine-frame window of 8th-order RASTA-PLP [5], energy, and delta-

RASTA-PLP features for every 25 ms window, stepped every 10 ms. The MLP is fully connected and has 153 inputs (9 frames with 17 features per frame), 1000 hidden units, and 56 outputs (one output for each phone[1]), and is trained using backpropagation with softmax normalization at the output layer. The system is trained on hand-transcribed phone labels (without embedded realignment). Using a multiple pronunciation lexicon (derived from the hand transcriptions), and a bigram language model, the word error rate (WERR) of this baseline system on the test set is 7.9%.

For our multi-band system, we divide the frequency range into four bands of [300-800Hz][2], [700-1600Hz], [1500-2700Hz], and [2100-3800Hz]. From the sub-bands, we derive [3rd, 3rd, 2nd, 2nd] order RASTA-PLP features, respectively, as well as energy and corresponding deltas. We train four MLPs on these acoustic features, that is, one on each sub-band. The input layer to each MLP has a context window of nine frames, for total input layer sizes of [72, 72, 54, 54] respectively. We choose hidden layer sizes of [497, 497, 372, 372], respectively, so that the total number of parameters in the four MLPs and the full-band system are roughly equal. There are 56 output units, one for every phone, as in the full-band MLP[1]. The frame-by-frame information from the four sub-band streams is combined using a *merger* MLP, which takes the output of the sub-band MLPs as input, has 300 hidden units, and an output of 56 phones[1]. The WERR on the test set for this merged multi-band system is 8.2%. The performance difference between the baseline and multi-band systems is not statistically significant.

3. IS PHONETIC INFORMATION LOST?

3.1. Experimental Setup

The first question we want to answer is whether any phonetic feature information is lost in multi-band ASR. For this analysis we use phone and broad category confusion matrices, as in the seminal studies of Miller and Nicely [7] on human speech recognition.

A confusion matrix (CM) is simply an extended matrix of *hits* and *misses* for all classes, as in Table 1. The column headings represent the features we intend to *transmit*, and the row headings correspond to the *received* features. In Table 1, for example, 93 instances of /s/ are perceived as /eh/. We use frame level phonetic classification on the test set for generating phone CMs. To better observe the patterns in the data, we collapse the phone CMs according to membership in broad category feature classes (as in Table 2), and generate feature confusion matrices (example in Table 3). We classify phonetic classes according to six broad categories: *CV* (consonant, vowel, silence), *duration* (short, long, mid), *frontness* (front, back, neither), *manner* (vowel, diphthong, liquid, glide, stop, closure, nasal, fricative, silence), *place* (high, low, mid, labial, dental, coronal, palatal, retroflex, velar, glottal, silence), and *voicing* (voiced, unvoiced).

To summarize the confusion matrix, we calculate mutual information (MI) for each CM [7] as $\sum_{i,j} p_{ij} log \frac{p_{ij}}{p_i p_j}$, where i is the feature we would like to transmit, and j is the feature that is perceived. We estimate the probabilities p_{ij}, p_i, and p_j from n_{ij}/n, n_i/n, and n_j/n, respectively, where n_i is the frequency of stimulus i, n_j is the frequency of response j, and n_{ij} is the frequency of the joint occurrence of stimulus i and response j in a sample of n observations.

We can further calculate the transmission of each phonetic sub-feature (e.g., sub-feature fricative $\in$ manner), by reducing the full CMs to a 2x2 CM for each $sub\_feature$ and $\overline{sub\_feature}$) (the results in Figure 3). Finally, the maximum possible feature transmission for the idealized condition is the MI of a matrix of the same dimensions and with the class priors on its diagonal.

[1]Note that some of the 56 phones do not occur in the NUMBERS database and have zero priors.

[2]Because we are testing on telephone quality speech, we disregard frequencies from 0 through 300Hz.

	t	s	eh	sil	...
t	5722	252	31	316	...
s	258	8495	110	1159	...
eh	11	93	3118	37	...
sil	436	2733	68	40237	...
...	...	...	...	...	...

Table 1: An example of a phone-based confusion matrix.

	vowel	consonant	silence
t	-	+	-
s	-	+	-
eh	+	-	-
sil	-	-	+
...	...	...	...

Table 2: An example of binary acoustic features for CV classification.

3.2. Observations

Figure 2 shows all features, and Figure 3 shows sub-features of *manner* transmitted as a percentage of the maximum. We observe the following:

1. Multi-band feature transmission is always as good as or better than the comparable full-band system, except for *frontness*. On average, 60.94% of the features are transmitted for the multi-band system compared to 59.06% for the full-band system for 54000 acoustic frames.

2. The results are consistent with our knowledge of acoustic phonetics, as, for example, we would expect the low frequency band to contain the most information about *voicing*. Comparing our results with [7], we observe similar patterns also for *fricatives* and *nasals*.

	vowel	consonant	silence
vowel	74393	6962	1816
consonant	6738	61030	5055
silence	2321	8922	49281

Table 3: An example of a feature-based confusion matrix.

3. Low and sometimes mid frequency bands (often band 1 and sometimes band 2) transmit most of the feature information alone. For example, band 2 transmits 87% of the *frontness* features that are transmitted by the full-band system.
4. There is much redundancy in phonetic information content in the sub-bands, as the sum of information transmission over all bands far exceeds 100%. Lippmann [6] has highlighted this redundancy as a source of human robustness to speech degradations.

   In the next section, we examine the transition asynchrony hypothesis.

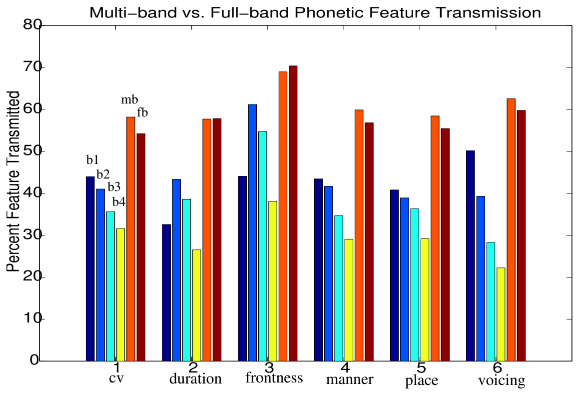

Figure 2: Phonetic features transmitted as a percentage of maximum possible, as measured by mutual information.

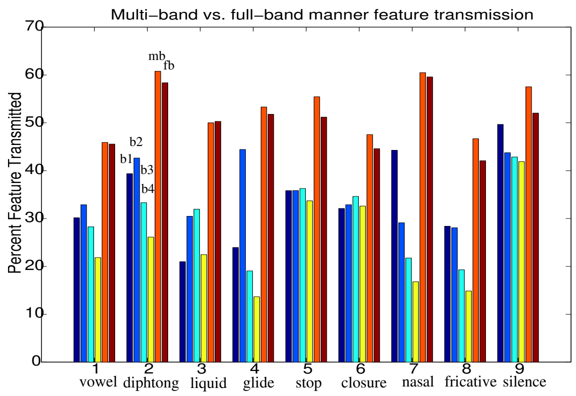

Figure 3: Manner of articulation features transmitted as a percentage of maximum possible, as measured by mutual information.

4. DO TRANSITIONS OCCUR ASYNCHRONOUSLY?

Multi-band researchers have posited that transitions occur asynchronously in sub-bands, and a phone or syllable level merging of multi-band streams may be necessary. In this section we study this hypothesis.

4.1. Experimental Setup

In order to obtain the phone transition boundaries, we perform forced alignment on each sub-band independently. Furthermore, to allow maximum freedom of shifting in transition boundaries, we perform embedded realignment (i.e., Viterbi realignment and retraining the MLP in each iteration) for six iterations. The WERR on the NUMBERS95 cross-validation set is our stopping criterion, and it reaches a minimum value after the second iteration of realignment.

Instead of using our usual multiple pronunciation lexicon, we use whole-sentence models in the forced alignment to insure that identical phone sequences are taken in each sub-band. We generate whole-sentence models using the phonetic hand-transcriptions and the corresponding average phone durations.

We also generate these statistics on the digitally-reverberated versions of the data, as well as on fast and slow speech. The reverberant data set was generated by convolving the clean set with an impulse response measured in a room having a reverberation time of 0.5 s and a direct-to-reverberant energy ratio of 0 dB. The cutoff for fast (slow) speech is set to one standard deviation above (below) the mean rate of the training set. The speaking rates were determined from a count of manually transcribed phones over non-silence regions.

For any given phone transition, we calculate the transition lags in each sub-band as compared to 1) the full-band, and 2) other sub-bands. Figure 4 show the histograms of average transition lags of the four sub-bands with respect to the full-band for broad phonetic categories, where each plot in row *feat1* and column *feat2* corresponds to a *feat1* $\longrightarrow$ *feat2* transition.

4.2. Observations

Examining the generated statistics, we observe that sub-band transitions do indeed occur asynchronously. More precisely:

1. Transition lags (with respect to the full-band transition boundaries) have a Gaussian distribution, with a mean close to zero, indicating that on average the transition lags happen in both directions, and a standard deviation of [2.8, 3.3, 5.0, 5.6] frames for the sub-bands, respectively. The higher the frequency range, the more shifted are the transition boundaries compared to the full-band.
2. More distant sub-bands have less agreement in transition boundaries, as the σ of transition lags between sub-bands 1 and 4 is 5.9 frames, and between sub-bands 1 and 2 is 3.8 frames.
3. 30% of the sub-band transitions do not occur within 50 ms of each other. Similarly 44%, 41%, and 21% of the transitions for reverberated, slow, and fast data, respectively, do not occur within 50 ms of each other.
4. Some broad category transitions are sharp (e.g., sil $\rightarrow$ stop), and some have a relatively flat distribution (e.g., vowel $\rightarrow$ liquid) (see Figure 4 for more examples).

For contrast conditions of speaking rate and room reverberation, we also found strong changes in transition timing, as reflected in a modified variance rather than a systematic difference in the means. Table 4 shows that for 3 out of the 4 bands, the standard deviation of the per-band lag decreases as speaking rate increases, which conforms to the intuition that phone durations decrease with rate. The table also suggests that the higher frequency transitions are most sensitive to speaking rate variations.

Table 4 further confirms our intuition that reverberation should affect transitions more at low frequencies than at high frequencies,

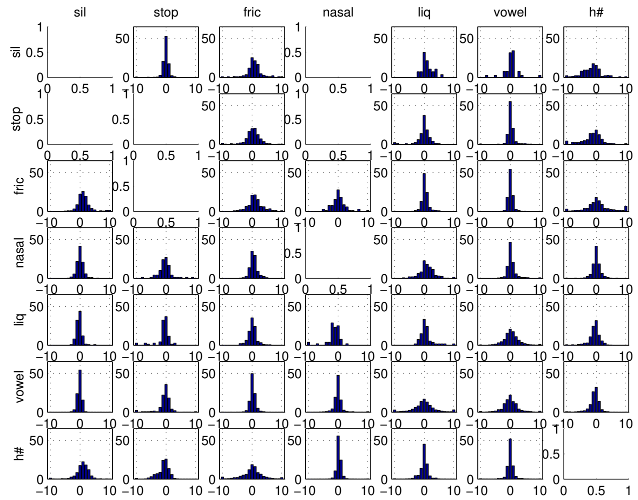

Figure 4: Histogram of average transition lags for broad phonetic categories for the four sub-bands. Each frame corresponds to 10 ms. /h#/ is the end/beginning of sentence silence.

Condition	band 1	band 2	band 3	band 4
Slow	3.7	3.6	9.8	9.2
Medium	2.8	3.1	4.2	5.1
Fast	2.1	4.1	2.8	3.6
Reverb	4.0	4.4	5.5	6.3
Clean	2.8	3.4	5.0	5.6

Table 4: Standard deviation for sub-band transition lags as compared to the full-band transition boundaries.

since most common room boundary materials are less absorptive at low frequencies, leading to longer reverberation times at those frequencies.

5. CONCLUSIONS

We have tested two common assumptions on multi-band ASR: 1) the objection of the critics of multi-band ASR that it is inherently inferior to a full-band approach because phonetic information is lost due to the division of the frequency space into sub-bands; and 2) the assumption by multi-band ASR researchers that transitions in bands often occur asynchronously (i.e., at different times than the full-band transition).

To study the first point, we calculated phonetic feature *transmission* for sub-bands. Not only did we fail to substantiate the above objection, but we observed the contrary. We confirmed the second hypothesis by analyzing the *transition* lags in each sub-band.

Our exploration of the first question further showed that, even when using a simple multi-band merging method, phonetic features are transmitted better (60.94% for our database) than the comparable full-band system (59.06%).

For the second question, we found that there is no consistent delay or expedition of phone transitions in a frequency-dependent manner, as the per-band transition lags had a mean close to zero. However, the spread of these transition lags were both dependent on frequency and on contrast conditions (speaking rate and reverberation). In particular, roughly one-third of the sub-band transitions in the control condition do not occur within 50 ms of each other. Furthermore, the high frequency band timings have a spread that is strongly dependent on speaking rate.

It appears that sub-band alignments can have significant timing deviations from the full-band alignments; thus, we would expect that there is a potential for improvements in acoustic modeling if longer time-scale information stream merging (i.e., phone or syllable) is used.

Acknowledgments

We would especially like to thank Brian Kingsbury for the CM script and for proof-reading services, and Eric Fosler-Lussier, Su-Lin Wu, and Dan Gildea for helpful discussions on lexicon creation and decoding. We acknowledge our colleagues Hervé Bourlard, Stéphane Dupont, Steve Greenberg, Hynek Hermansky, and Sangita Tibrewala for multi-band collaboration over the past two years. We thank Jim West and Gary Elko, from Bell Labs, and Carlos Avendano, now at the University of California, Davis, for collecting the room impulse responses and making them available to us. This work was supported by Mentored Research Fellowship by the University of California, European Community Basic Research grant (Project Sprach), and the International Computer Science Institute.

6. REFERENCES

[1] Jont B. Allen. How do humans process and recognize speech? *IEEE Trans. on Speech and Audio Proc.*, 2(4):567–577, Oct 1994.

[2] Hervé Bourlard and Stéphane Dupont. Subband-based speech recognition. In *ICASSP*, volume II, pages 125–128, May 1997.

[3] Hervé Bourlard and Nelson Morgan. *Connectionist Speech Recognition – A Hybrid Approach*. Kluwer Academic Press, 1994.

[4] Harvey Fletcher. *Speech and Hearing in Communication*. Krieger, New York, 1953.

[5] Hynek Hermansky and Nelson Morgan. RASTA processing of speech. *IEEE Transactions on Speech and Audio Processing*, 2(4):578–589, Oct 1994.

[6] Richard P. Lippmann. Speech recognition by machines and humans. *Speech Communication*, 22(1):1–15, 1997.

[7] George A. Miller and Patricia E. Nicely. An analysis of perceptual confusions among some English consonants. *JASA*, 27(2):338–352, Mar 1955.

[8] Nikki Mirghafori. An alternative approach to automatic speech recognition using sub-band linguistic categories. Thesis Proposal (http://www.icsi.berkeley.edu/~nikki/papers/thesis_prop.ps), Dec 1996.

[9] Sudhakar Rao and Wiliam A. Pearlman. Analysis of linear prediction, coding, and spectral estimation from subbands. *IEEE Transactions on Information Theory*, 42(4):1160–1178, Jul 1996.

[10] Sangita Tibrewala and Hynek Hermansky. Sub-band based recognition of noisy speech. In *ICASSP*, volume II, pages 1255–1258, May 1997.