DETECTING BANDLIMITED AUDIO IN BROADCAST TELEVISION SHOWS

Mark C. Fuhs, Qin Jin, and Tanja Schultz

InterACT, Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213
{fuhs,qjin,tanja}@cs.cmu.edu

ABSTRACT

For TV and radio shows containing narrowband speech, Speech-to-text (STT) accuracy on the narrowband audio can be improved by using an acoustic model trained on acoustically matched data. To selectively apply it, one must first be able to accurately detect which audio segments are narrowband. The present paper explores two different bandwidth classification approaches: a traditional Gaussian mixture model (GMM) approach and a spline-based classifier that categorizes audio segments based on their power spectra. We focus on shows found in the DARPA GALE Mandarin training and test sets, where the ratio of wideband to narrowband shows is very large. In this setting, the spline-based classifier reduces the number of misclassified wideband segments by up to 95% relative to the GMM-based classifier for the same number of misclassified narrowband segments.

Index Terms— Speech processing, speech recognition, pattern classification, telephony

1. INTRODUCTION

One of the challenges presented when performing STT on television and radio shows is that narrowband audio can be embedded within the show. This typically occurs for particular speakers who participate via telephone, which is typically 8kHz, 8 bit μ-law encoded and band-pass filtered to between 300Hz and 3.6kHz. Using acoustic models matched to the test condition is well known to improve STT accuracy; however, to use a narrow-band acoustic model, one must first determine to which audio segments the narrow-band model should be applied.

Detection can be particularly challenging in broadcast shows because distortion in the upper bands is frequently present to varying degrees and is correlated with the narrowband speech; Figure 1 shows three examples. Background music and other wide-band audio can also be found mixed in during the narrow-band speech. For the types of shows addressed by DARPA's GALE program, narrowband segments are a small minority, making it critical to minimize the false-positive rate of detecting a narrowband segment, since using a narrowband acoustic model to decode wideband audio typically degrades STT performance as much as using it to decode narrowband audio will improve performance.

Gaussian mixture models (GMMs) are a popular statistical framework for detecting narrowband audio [1, 2, 3]. We compare the GMM approach with an approach based on classifying the audio power spectrum. Section two describes the training and testing conditions, as well as a description of the wide-band and narrow-band acoustic models. Sections three and four describe the GMM and spectrum-based classifiers, respectively. Conclusions are drawn in section five.

2. TRAINING AND TESTING CONDITIONS

The present experiments focus on Mandarin TV shows provided within DARPA's GALE program. Unfortunately, references do not include the bandwidth of the audio segments. For training, speakers were selected from the GALE training data releases (excluding P3R1), and one segment of each speaker was hand-labeled. All segments from that speaker were then included as training data with the bandwidth label assigned from the single hand-labeled segment (15 hrs total). Speaker selection was random, but biased by a primitive classifier to exclude speakers that were "obviously" wide-band speakers in order to achieve a better training set balance between bandwidth classes.

The test set comprised the entire P3R1 GALE data release, excluding narrow-band shows from the Phoenix network (see below), which is 303 hrs of audio. A large test set was chosen since narrowband audio is relatively uncommon in shows in this corpus (0.4% of segments). (By contrast, call-in radio shows, for example, would have a much higher ratio.) The test set was labeled semi-manually, using the spectrum-based classifier (see Section 4) run on multiple segments per speaker concatenated together – at least 15 sec of audio. All segments classified as narrowband and segments classified as wideband that were close to the decision surface were then hand-labeled. As Table 1 shows, the likelihood of a narrowband segment decreased rapidly with distance from the decision surface, suggesting that, without manually labeling ev-

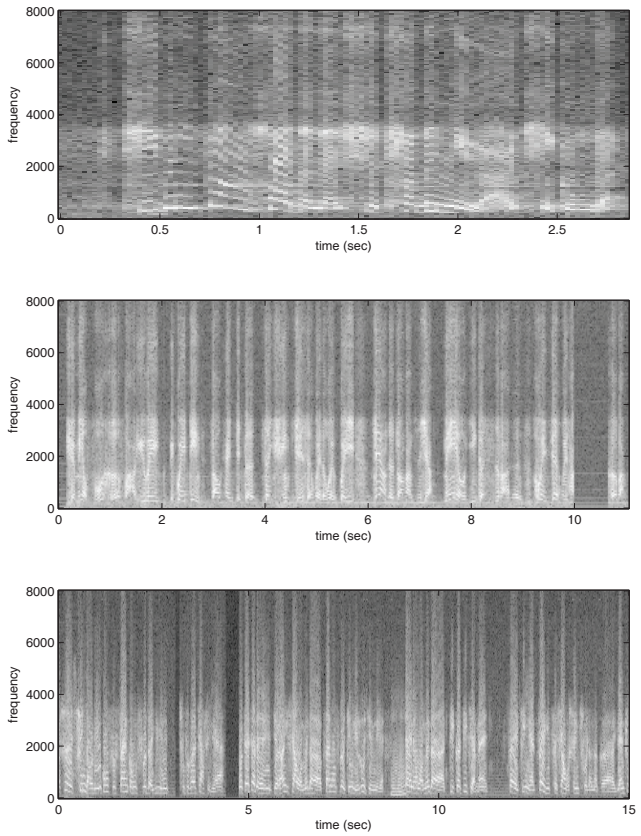

Fig. 1. Three examples of telephone audio segments found within TV show audio otherwise sampled at 16kHz. Significant distortion is present above 4kHz.

ery segment, the vast majority of narrowband segments have been identified. Moreover, narrowband segments in the 1.25 to 1.5 distance category were subjectively often quite difficult to identify manually due to high levels of distortion.

To demonstrate that STT performance benefits from identifying the narrowband segments, narrowband segments were decoded using "wideband" and "narrowband" SI acoustic models. Briefly, both models were trained on the Y1Q2, Y1Q4, Y1 interim and P2R2 GALE releases from the LDC, approximately 840 hrs in total. These training corpora included shows from several networks, though most were from CCTV, with audio bandwidth extending to the 7-8kHz range, and Phoenix, with audio power extending only to 4.7kHz. The "wideband" model was trained on this mixed set of shows as they appear in the training corpus. The "narrowband" acoustic models were originally build to improve performance on the Phoenix shows: all non-Phoenix shows were low-pass filtered to 4.7kHz. Nonetheless, STT performance on telephone data is significantly better using this latter model. Both acoustic models use initial-final phonetic models, 7000 clustered states, and up to 32 Gaussians per state. A full system description of our Mandarin STT system

Distance to decision surface	Segments	% narrowband
0 to 0.5	102	18.6%
0.5 to 1.0	420	4.5%
1.0 to 1.25	358	2.0%
1.25 to 1.5	552	0.9%

Table 1. Semi-manual labeling of the test set. As the distance from the SVM decision surface grows, the likelihood of an incorrectly classified segment decreases precipitously.

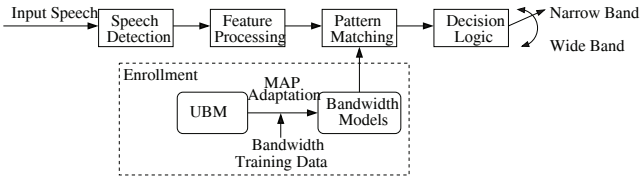

Fig. 2. Schematic overview of the GMM-based bandwidth classification system.

can be found in [4]. On the GALE dev07 test set, WERs were 18.5% for the wideband model and 20.9% for the narrowband model.

3. GMM-BASED BANDWIDTH ESTIMATION

Gaussian mixture models were used as generative models for a two-class (narrowband / wideband) classification problem. Our GMM-based bandwidth classification system, shown in Figure 2, consists of five key components: speech detection (or silence removal), feature processing, pattern matching, decision logic, and enrollment. Speech detection based on the energy of the speech signal is applied to remove silence before further processing. For the feature processing we apply 13-dimensional Mel-frequency Cepstral Coefficient (MFCC). Cepstral Mean Normalization (CMN) is applied over MFCC features to remove channel effects. Using these features the pattern matching component evaluates them under stored generative models and calculates a probability for each model. The resulting scores are fed into the decision maker, where the system finally decides on the bandwidth category of the input speech. The decision threshold was defined by whether the log likelihood ratio (LLR) of the segment under each model was above or below a fixed constant.

Generative models must first be trained for each bandwidth, a process commonly referred to as enrollment. The bandwidth models are trained by Maximum A Posteriori (MAP) adapting on a Universal Background Model (UBM). In our system, the UBM representing a general bandwidth space is trained with 1024 Gaussian mixtures using both narrowband and wideband data.

Figure 3 shows a DET curve of error rates for the GMM-based classifier. Based on the percentage of misclassified segments, the GMM-based classifier is capable of detecting

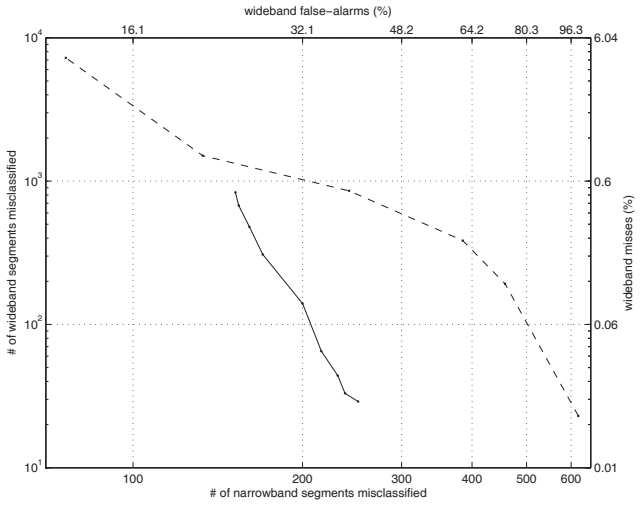

Fig. 3. Plot of raw number of errors (bottom/left axes) and corresponding error rates (top/right axes) for the two classifiers. Dashed line: GMM-based classifier using different LLR decision thresholds. Solid line: Spline-based classifier using different weightings of the training set classes.

87.5% of the narrowband segments with only a 4.4% false-positive rate. However, the high ratio of wideband to narrowband segments in the test set also means that there will be 93 times as many misclassified wideband segments as narrowband segments. Other decision thresholds improve the error ratio, but there is no point at which the benefit from applying the narrowband acoustic model to the correctly classified narrowband segments would compensate for the harm of applying it to the incorrectly classified wideband segments.

4. SPLINE-BASED BANDWIDTH ESTIMATION

4.1. Power spectrum splines

Though significant distortion above 4kHz can be found in telephone speech contained within TV and radio shows, there is typically some difference in power between the speech signal below 4kHz and the distortion above. Moreover, there tends to be a recognizable drop in the power spectrum in the 3.5-4kHz range. This second approach attempts to identify narrowband segments based on the presence of such a drop.

To identify this drop, a spectrogram of each utterance, based on 50ms frames, was summed over the frames to produce an average power spectrum, $p(f)$. This power spectrum was fit with two splines:

$$\begin{aligned} p_{\text{lin}}(f) &= af + b \\ p_{\text{nonlin}}(f) &= a_1 f + b + c\tanh(d(f+m)) \\ &\quad + a_2 (f - f_{5\%}) I(f > f_{5\%}) \end{aligned}$$

The p_{lin} spline is a simple linear function. The p_{nonlin} spline is composed of a linear function, $a_1 f + b$, plus a non-linear sigmoidal function, $c\tanh(d(f+m))$, intended to represent an abrupt drop in power around a particular frequency. After the drop is nearly complete – the sigmoid is 5% from its lowest value – the linear portion of the spline continues. Its slope may be adjusted to be different from the slope prior to the sigmoid by the term $a_2 (f - f_{95\%}) I(f > f_{95\%})$, where I is the indicator function. These splines were motivated by the observation of a large number of wide- and narrowband audio segments and can be thought of as generative models of the power spectrum of each segment class.

The p_{lin} spline was regressed to the observed power spectrum by the standard least-squares approach. The non-linearity in the p_{nonlin} spline required that it be fit with gradient-descent. A two-pass approach was used, first fitting the spline to the observed power spectrum from 1.5kHz to 6.75kHz. The first pass gave an estimate of the beginning of the drop, $f_{95\%}$, and the end of the drop, $f_{5\%}$. During the second pass, only the observed spectrum from $f_{95\%}$ - 1.25kHz to $f_{5\%}$+2.25kHz was used in order to minimize the effects of power variations near the higher and lower frequencies of the spectrum. Examples are shown in Figure 4.

4.2. SVM classifier

Once the two splines were fit to the power spectrum of an audio segment, the following features were extracted:

- The coefficient of determination of the nonlinear spline relative to the linear spline. This is the proportion of the p_{lin}residual that is removed by using the p_{nonlin} model instead.
- The "differential magnitude" of the drop: over the frequency span occupied by the non-linearity in p_{nonlin}, how much more of a power drop is measured according to p_{nonlin} than p_{lin}.
- The power at the frequency where 95% of the drop is completed.
- Coefficients c and d in p_{nonlin}.
- The average power over the entire show.

These features were scaled and used as input to an SVM with a cubic polynomial kernel. Kernel parameters and the slack variable weight were tuned using 20-way cross-validation on the training set.

Figure 3 shows a DET curve of error rates for the spline-based classifier. Compared to the GMM-based classifier, the spline-based classifier shows substantially lower wideband misclassifications for the same number of narrowband misclassifications. Narrowband misclassifications were generally because the power spectrum showed only a small drop

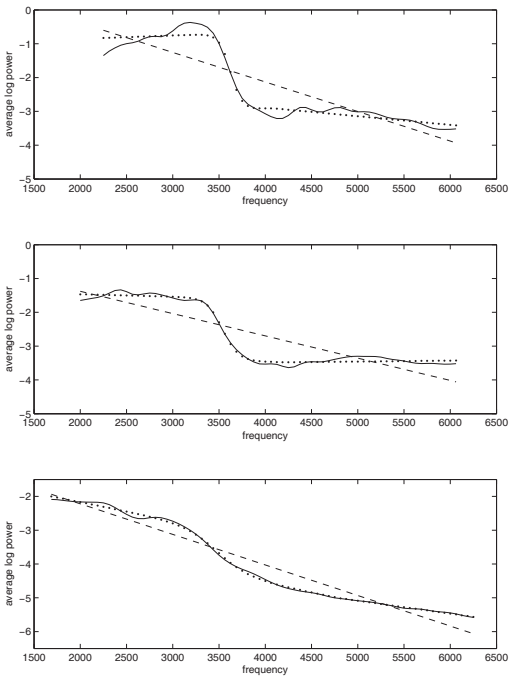

Fig. 4. Power spectra and splines for the three examples shown in Figure 1. The solid line indicates the observed power spectrum. The dashed line is the p_{lin} spline and the dotted line is the p_{nonlin}. While the spline-based approach is very good at detecting discrete power drops, even with significant upper-band distortion (top, middle), more gradual power drops (bottom) result in power spectra that are more difficult to differentiate from a wideband signal. Though shifted, the y-axis scales are the same for visual comparison.

in power relative to a more linear decline, as in Figure 4 (bottom). However, the ability of the spline-based classifier to trade off wideband and narrowband misclassifications is more limited, and may not be applicable in corpora where narrowband segments with high distortion are very common. One potential benefit of the spline-based approach is that the features derived from the splines are largely independent of the cutoff frequency. Thus, the same classifier can recognize band-limited audio with different cutoff frequencies without retraining.

To test the usefulness of the spline-based classification on STT performance, we decoded the set of segments classified as narrowband either manually or by the spline-based classifier using the acoustic models described in Section 2. Table 2 presents the results using an SVM class weight optimized on the GALE dev07 test set. The WER of this set of segments is significantly improved by identifying and decoding them using a narrowband acoustic model. The spline-based classification, which identifies 65% of the segments, shows 71% of the benefit that manual classification could achieve.

Condition	# wide	# narrow	WER
Wideband model only	656	0	41.9
Spline-based classifier	254	402	40.4
Manual classification	33	623	39.8

Table 2. Benefits of using the spline-based classifier on STT WER. Listed are the number of segments decoded using each acoustic model and the combined WER of all 656 segments.

Interestingly, the 33 segments that were identified manually as wideband but by the classifier as narrowband had very poor WERs (typically over 50%) using either acoustic model. Decoding the 656 segments using the narrowband acoustic model also yielded 39.8%.

5. CONCLUSIONS

We present results from two very different approaches to detecting audio bandwidth. In a setting where the ratio of wideband to narrowband segments is very large, the GMM-based classifier misclassified too many wideband segments for the classifier to be beneficial. The spline-based classifier was specifically tailored to detect the abrupt drop in power typical of narrowband segments embedded in wideband audio. For a given number of narrowband misclassifications, the spline-based classifier could achieve as much as a 95% reduction in wideband misclassifications relative to the GMM-based classifier, enabling the selective application of a narrowband acoustic model to improve STT accuracy.

6. ACKNOWLEDGMENTS

This work is partly supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR0011- 06-2-0001. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DARPA.

7. REFERENCES

[1] T. Hain, S. Johnson, A. Tuerk, P. Woodland, and Young S., "Segment generation and clustering in the htk broadcast news transcription system," in *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, 1998.

[2] D. A. Reynolds and P. Torres-Carrasquillo, "Approaches and applications of audio diarization," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2005.

[3] X. Zhu, C. Barras, S. Meignier, and J.-L. Gauvain, "Combining speaker identification and BIC for speaker diarization," in *Proc. Interspeech*, 2005.

[4] R. Hsiao, M. Fuhs, Y.-C. Tam, Q. Jin, and T. Schultz, "The CMU-InterACT 2008 Mandarin transcription system," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2008.