MODEL-BASED MONAURAL SOURCE SEPARATION USING A VECTOR-QUANTIZED PHASE-VOCODER REPRESENTATION

Daniel P. W. Ellis and Ron J. Weiss

LabROSA, Dept. of Elec. Eng.
Columbia University
New York NY 10027 USA
{dpwe,ronw}@ee.columbia.edu

ABSTRACT

A vector quantizer (VQ) trained on short-time frames of a particular source can form an accurate non-parametric model of that source. This principle has been used in several previous source separation and enhancement schemes as a basis for filtering the original mixture. In this paper, we propose the "projection" of a corrupted target signal onto the constrained space represented by the model as a viable model for source separation. We investigate some parameters of VQ encoding, including a more perceptually-motivated distance measure, and an encoding of phase derivatives that supports reconstruction directly from quantizer output alone. For the problem of separating speech from noise, we highlight some problems with this approach, including the need for sequential constraints (which we introduce with a simple hidden Markov model), and choices for choosing the best quantization for overlapping sources.

1. INTRODUCTION

Separating multiple, overlapping acoustic sources given only a few sensors is an underdetermined problem that requires additional constraints to be solved. We are particularly interested in the single-channel case where no spatial information is available, yet informally it is still possible to 'hear out' individual components. The only remaining constraints are the limitations on the possible forms of the source signals themselves; in practice, real-world sound sources of interest have structured and constrained properties (such as the stationarity, periodicity, and limited spectral variation of speech), and the task of monaural source separation can be viewed as a problem of suitably capturing and applying these constraints.

Our work is inspired by the single-microphone separation system described by Roweis [1, 2]. In the MAXVQ system, large vector-quantization (VQ) codebooks are trained on particular sources of interest (such as the voice of a particular speaker) to capture the constrained set of waveform snippets that constitute that source's 'palette'. VQ is applied on the short-time Fourier transform (STFT) magnitude – i.e. columns of a narrowband spectrogram – to capture local stationarity of the source spectrum, and to hide variability arising from the arbitrary alignment between analysis window and waveform (which will mainly appear in the STFT phase). Separation of single or multiple sources can then be achieved by finding the best-matching codewords consistent with an observed mixture. However, the resulting quantized representation, lacking phase, is not directly invertible. Instead, estimated magnitudes from the inference are used as a basis for filtering the original mixture (e.g. a time-varying Wiener filter, or a more extreme binary time-frequency mask) to recover an estimate of the original target source. Related approaches have been investigated by several other researchers, including [3], who derive soft masks from the posterior probabilities of each cell belonging to a particular source, [4], who learn separate but coupled models for multiple frequency subbands, and [5] who infer distributions over the target speech magnitudes.

By contrast, we are interested in building a model of the original signal that is sufficient to permit a perceptually satisfactory resynthesis from the representation alone. This has two implications: first, the set of signal elements covered by the codebook will need to be very large (and spaced according to a perceptual metric) in order to allow high-quality reconstruction of the original signal; and secondly, we must have a mechanism for reconstructing phase values for decoded frames, for instance by including phase-related information along with each codebook entry.

Given a model that supports a perceptually-adequate reconstruction, the problem of source signal separation may be reduced to inference or estimation of the model parameters corresponding to one or more individual sources within a mixture. In the limit, a VQ system will encode all the possible parts of feature space in which the sound of the modeled source may reside, and exclude all others. For a source with a limited space of possible short-time sound frames, but spread over a wide range of feature space, it may be possible to remove noise and other interference simply by projecting the observed, interfered feature frames onto the best-matching point in the subspace spanned by the source model.

In this work, we are concerned with the human voice. The voice of any particular individual is limited in its usual range, and thus there is hope that these limits may provide a basis for inference and separation. We are motivated by the hope that current computational power and available databases make possible simple, nonparametric models such as VQ, which are able to adequately and accurately capture the limited subspace of complex sources, meaning that separation can be achieved simply by projecting a mixture onto this limited space. This is the same idea as in [6], although here we do not use spatial information nor such a large codebook, and we have a different approach to transitions.

In the next section, we describe some empirical experiments with building large VQ codebooks for a single speaker's voice. In

Joint first authors are listed in alphabetical order.

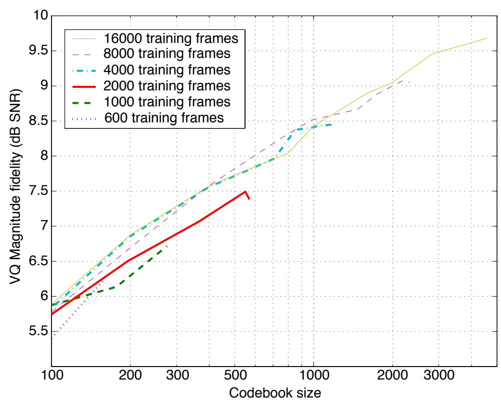

Fig. 1. VQ fidelity (as SNR of STFT magnitudes) on a held-out test as a function of codebook size (horizontal axis) and training set (different traces).

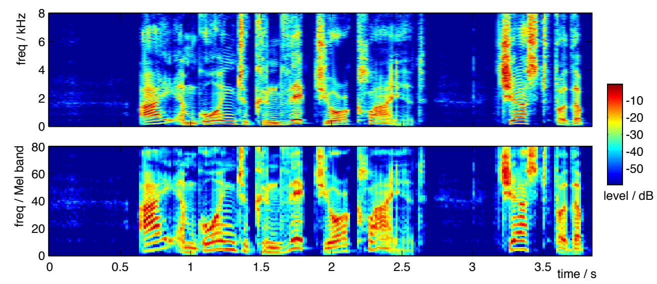

Fig. 2. Comparison of spectrograms using a linear frequency axis (i.e. the direct STFT magnitude), and, below, the same fragment displayed on the 80-bin Mel axis.

section 3, we describe how we used the phase vocoder representation as a basis for invertible short-time spectral codewords, and compare it to iterative phase reconstruction. Section 4 discusses the use of such models to recover signals from mixtures, including Markov modeling to capture temporal constraints. Finally, we conclude with a discussion of the limitations of the current approach.

2. VECTOR QUANTIZATION

Since we were interested in learning large models of a single source, we used an audiobook CD providing us with several hours of speech from a single male actor, recorded in consistent, clean conditions. We pre-emphasized the voice with FIR filter $1-0.95z^{-1}$. Our first experiment was to establish the variation of VQ distortion with the size of the codebook and training set. To capture harmonic structure, we use an STFT analysis window of 32 ms (512 samples at 16 kHz sampling rate), and an 8 ms hop between frames to ease overlap-add (OLA) reconstruction.

Figure 1 shows the fidelity of VQ quantization as a function of training set and codebook size. VQ codebooks were trained by conventional k-means clustering initialized with random samples from the data [7]. (Throughout this paper, we report performance in terms of signal-to-noise or signal-to-distortion ratio which we refer to as "magnitude SNR", calculated as the mean-squared error (MSE) in the STFT magnitude domain rather than on the waveform, since our primary quantization is in this domain.)

Even for our largest codebooks, fidelity is quite poor. It appears to be improving by about 0.5 dB for each doubling in codebook size, predicting that to achieve 20 dB SNR on this full-resolution spectrum would require a codebook of around 8 billion entries! Thus our hope that current computational power makes it easy to describe a source as rich as the human voice with a nonparametric model, appears overly optimistic.

This VQ approach seeks to optimize the mean-squared error under the distance measure used to quantize points to codewords within the k-means clustering. Two problems with mean-squared error in the linear STFT magnitude domain are (a) overemphasis of the high frequency bands, since the constant bandwidth bins of the STFT are a poor match to perceptual sensitivity to spectral detail which follows a more constant-Q behavior; and (b) inappropriate balance between errors in low-amplitude and high-amplitude values: perceptual systems tend to perceive errors in proportion to the underlying value, so a fixed difference becomes less salient as the absolute level of the quantities being compared increases.

To create a codebook that comes a little closer to uniform perceptual distribution, we modified the distance measure to work on cube-root-compressed, Mel-scale frequency vectors. 257-bin STFT spectral magnitude vectors covering 0-8 kHz are mapped into 80 Mel-frequency bins [8]. Each Mel bin is integrated over a triangular window in linear frequency spanning the interval between the center frequencies of the neighboring bands; for an 80-bin Mel spectrum, this gives a frequency resolution of approximately 40 Hz in the constant-bandwidth low-frequency region – fine enough to retain harmonic detail. Figure 2 compares linear- and Mel-frequency spectrograms. We use Mel-scale spectrograms for subsequent figures.

To address the issue of nonlinearity in sensitivity to errors, we compressed the linear magnitudes by using the common cube-root approximation to perceived loudness. While the 80-bin, cube-root compressed spectra are used to build the clusters, the codebook entries are the centroids of the clusters in the original linear STFT magnitude space. Figure 3 compares 800-entry codebooks constructed using standard and perceptually-modified distance metrics. We expected that the Mel-frequency warping would cause more codewords to be used for frames with energy and structure in lower frequencies, but in this example the two codebooks look very much alike. In some cases, we felt that the perceptual codebooks gave more pleasant reconstructions.

3. PHASE-VOCODER REPRESENTATION

To reconstruct sound with OLA resynthesis, it is necessary to have both the magnitude and phase of the STFT. In the previous systems, the phase is taken from the mixture STFT; where the mixture is dominated by the target source, this is a good estimate, but it is less appropriate elsewhere.

We are interested in defining a model that permits reconstruction based on the model representation alone. Since it depends on the arbitrary alignment of the time frame, phase is not stable even for a signal with stationary spectral content, and is thus not amenable to direct quantization. The *phase vocoder* [9, 10] takes advantage of the relative unimportance of absolute phase and the fact that the *frequency* of a sinusoidal component gives the rate of phase *change* between adjacent frames to extract a stable representation as the phase *derivative* along time, or instantaneous fre-

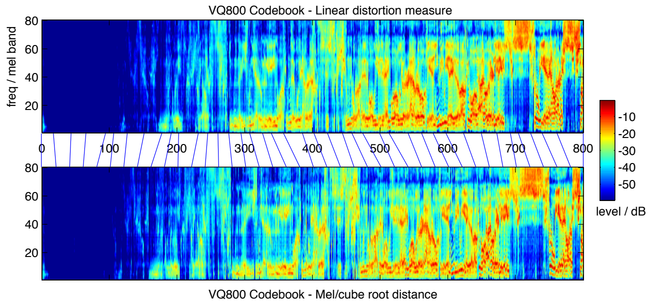

Fig. 3. Complete 800-entry codebooks designed using standard and preceptual distortion meaures. The codebooks are sorted to place similar codewords nearby, and to present comparable ordering. Lines connect most-similar codewords between the codebooks.

quency (IF). If a number of spectral frames share a strong peak at a particular frequency, we expect the phase derivative to be similar for all the frames, and thus we expect IF to quantize well in our codebooks, even if it is not used in the clustering distance metric. Resynthesis consists of reading the magnitudes and IFs from a sequence of codewords, then integrating the phase-derivatives within each frequency channel to reconstruct a smooth, consistent phase function suitable for OLA resynthesis. To reduce "phasiness" [10], we also quantize and store across-frequency phase differences, and move cumulated phases towards these on reconstruction.

A second approach to reconstructing a full signal from magnitudes alone is to iteratively perform full inversion, starting from some random phase, then re-analyze the OLA result by the STFT, replace the magnitudes (which in general will have been distorted by constructive and destructive interference between frames) with the desired magnitudes while retaining the new phase, perform OLA reconstruction again, and iterate until the analysis magnitude is suitably close to the intended values [11, 12]. We found this approach to work well, as illustrated in figure 4, which shows the reconstruction magnitude SNR for different starting points, and as a function of iteration number. In this case, the magnitude SNR is based on the re-analysis of the reconstructed waveform, so it will include the magnitude distortion arising from OLA phase interactions. We see that quantizing IF significantly improves the starting point. Convergence to the intended magnitudes (indicated by the upper line) is quite slow.

4. MODEL-BASED SEPARATION

Returning to our original motivation for looking at nonparametric signal models, we want to separate the modeled source from interference by 'projecting' the mixture onto the constrained space represented by the model; as the applied constraints become stricter, and as the interference increases in difference from the target source, this approach should become more successful. As a simple illustration, we looked at separating speech from speech-shaped noise at various SNRs. Separation consists of estimating, at each time step, the codeword best matching the target speech in the mixture, then resynthesizing from these codewords alone. Finding the best codeword at each time step can be done several ways: we could simply quantize the mixture spectrum with the source-specific VQ, and hope that the reconstruction (which will at least

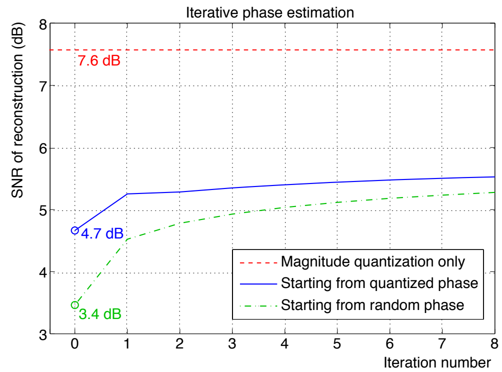

Fig. 4. Magnitude SNR accuracy for different approaches to phase reconstruction. Top trace shows upper limit imposed by magnitude quantization; middle trace shows phase integrated from quantized instantaneous frequency, through successive iterations of phase reestimation; bottom trace starts from random phase (from STFT analysis of equivalent-duration white noise).

lie within the subspace of the target source) is a fair approximation of the target signal – equivalent to assuming that the interference has contributed Gaussian noise (of constant variance) to every spectral magnitude value.

A more sophisticated approach is to recognize that signal components are more likely to be obscured if they are lower in energy, and that the distribution of level of corrupted magnitudes will be skewed towards the positive side – since the only case in which additive corruption results in reducing the STFT magnitude of a cell is if the corruption is of about the same energy and at a canceling phase. In most cases, the noise energy in a cell will likely be much larger or smaller than the target, and it is increasingly likely to be much larger when the target magnitudes are small. Thus, some measure of the posterior likelihood of each codeword that asymmetrically penalizes cells in which the magnitude is smaller than expected, and that gives more weight to the match for the largest codeword dimensions, should make better inferences. Given estimates or assumptions for the distribution of noise energies, the Bayes-optimal codeword can be selected.

Figure 5 gives an example of reconstructing noisy speech with the learned codebook. The top two panes show the original speech example before and after adding noise filtered to have the same average spectrum and matching energy (0 dB SNR, resulting in a magnitude SNR of 1.8 dB). The third pane shows the result of directly quantizing this mixture with a 400-entry codebook. The constraints imposed by the codebook are sufficient to remove noise in certain frames, particularly when there is significant high-frequency energy (e.g. at $t = 1.3, 2.3s$) which can only be matched by imposing low energy in the lower bands. However, during lower energy regions, the codebook is able to find codewords approximately matching the noise spectrum, so the pauses in the speech are filled with noise-like codewords and the magnitude SNR, at -0.6 dB, is actually worse than the original mixture.

A more complete model of the source will also exploit the temporal constraints exhibited by the source. To capture this, we trained a discrete-observation hidden Markov model on top of the VQ. In theory, we could learn transition probabilities between every pair of codewords, but getting good estimates for this N^2-sized matrix (160,000 entries in this small example) requires prohibitive

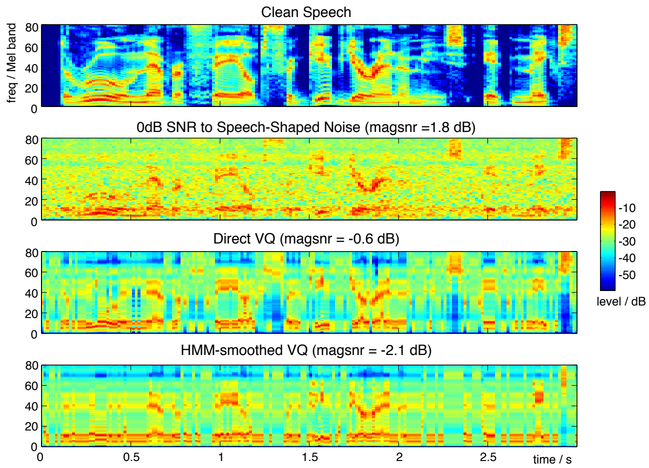

Fig. 5. Mel-scale spectrograms of clean speech (top), mixed with speech-shaped noise at 0 dB SNR (next), then 400-entry codebook reconstructions with and without smoothing by a 50-state HMM (lower two panes).

amounts of training data. Instead, we build a Markov model with a smaller space of 50 states, then learn both the reasonably-sized transition matrix and a discrete distribution of codewords emitted by each state. In this way, the detail of spectral structure can still be defined by a large codebook, but the broad dynamics (of staying within a particular neighborhood, or transitioning to a different area) can be tractably learned and represented. The bottom panel shows the codeword sequence inferred by this model. We can see some examples of improved continuity, for instance in the formant transitions at $t = 1.0, 1.8s$. However, another effect is to fill in more noise-matching frames in the silent gaps, and to delete some of the ranges where high-frequency bursts were well captured, leading to an overall significant worsening of magnitude SNR to -2.1 dB.

5. DISCUSSION AND CONCLUSIONS

The current results leave much to be desired. Firstly, the VQs presented cannot provide good SNR even in ideal circumstances; instead, we are looking at factoring the representation into two or more independent codebooks that can then cover exponential signal spaces. We are investigating independent component analysis (ICA) and co-operative VQ [13] for this.

The second problem is that the models are able to find reasonable fits to interference, rather than excluding it by the projection. Asymmetric functions based on numerical evaluations of the likelihood of different codewords under loosely-specified noise conditions should help this, although we wish to avoid constructing an on-line noise model (which is vulnerable to nonstationary noise).

Even with these problems solved, this model is still specific to a particular speaker at a particular level, etc. This, however, may be amenable to model adaptation techniques that, by setting a small number of warping parameters, result in a reasonably precise model for any of a range of speakers. Factorizing the codebook, and adapting each factor separately, may help further.

In conclusion, the purpose of this paper was to present non-parametric models as a viable tool for source separation. While this case is not irresistibly established, we believe the general framework is clear and, given that improving computational resources work in favor of this approach, we are still confident that this approach will yield exciting results.

6. ACKNOWLEDGMENTS

This work is supported by the Columbia Academic Quality Fund and the National Science Foundation (NSF) under Grant No. IIS-0238301. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF.

7. REFERENCES

[1] S. Roweis, "One-microphone source separation," in *Advances in NIPS 11*, pp. 609–616. MIT Press, Cambridge MA, 2001.

[2] S. Roweis, "Factorial models and refiltering for speech separation and denoising," in *Proc. EuroSpeech*, Geneva, 2003.

[3] A. M. Reddy and B. Raj, "Soft mask estimation for single channel speaker separation," in *Proc. ISCA Tutorial and Research Workshop on Statistical and Perceptual Audio Processing SAPA-04*, Jeju, Korea, October 2004.

[4] M. J. Reyes-Gómez, D. P. W. Ellis, and N. Jojic, "Multiband audio modeling for single channel acoustic source separation," in *Proc. IEEE Int. Conf. Acous., Speech, and Sig. Proc.*, Montreal, 2004, pp. V–641–644.

[5] T. Kristjansson, H. Attias, and J. Hershey, "Single microphone source separation using high resolution signal reconstruction," in *Proc. IEEE Int. Conf. Acous., Speech, and Sig. Proc.*, Montreal, 2004, pp. II–817–820.

[6] J. Nix, M. Kleinschmidt, and V. Hohmann, "Computational auditory scene analysis by using statistics of high-dimensional speech dynamics and sound source direction," in *Proc. Eurospeech*, Geneva, 2003, pp. 1441–1444.

[7] R. M. Gray, "Vector quantization," *IEEE ASSP Magazine*, vol. 1, no. 2, pp. 4–29, 1984.

[8] D. P. W. Ellis, "PLP and RASTA (and MFCC, and inversion) in Matlab," `http://www.ee.columbia.edu/~dpwe/resources/matlab/rastamat/`.

[9] J. L. Flanagan and R. M. Golden, "Phase vocoder," *Bell System Technical Journal*, pp. 1493–1509, November 1966.

[10] J. Laroche and M. Dolson, "Improved phase vocoder time-scale modification of audio," *IEEE Trans. Speech and Audio Proc.*, vol. 7, no. 3, pp. 223–232, May 1999.

[11] M. Slaney, D. Naar, and R. F. Lyon, "Auditory model inversion for sound separation," in *Proc. IEEE Int. Conf. Acous., Speech, and Sig. Proc.*, Adelaide, 1994, pp. II–817–820.

[12] D. Griffin and J. Lim, "Signal estimation from modified short-time fourier transform," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 32, pp. 236–242, 1984.

[13] Z. Ghahramani, "Factorial learning and the EM algorithm," in *Advances in Neural Information Processing Systems (NIPS) 7*, G. Tesauro, D. S. Touretzky, and T. K. Leen, Eds., 1995, pp. 617–624.