SABR: Sparse, Anchor-Based Representation of the Speech Signal

Christopher Liberatore, Sandesh Aryal, Zelun Wang, Seth Polsley, Ricardo Gutierrez-Osuna

Department of Computer Science and Engineering, Texas A&M University, United States

{cliberatore, saryal, wang14359, spolsley, rgutier}@tamu.edu

Abstract

We present SABR (Sparse, Anchor-Based Representation), an analysis technique to decompose the speech signal into speaker-dependent and speaker-independent components. Given a collection of utterances for a particular speaker, SABR uses the centroid for each phoneme as an acoustic "anchor," then applies Lasso regularization to represent each speech frame as a sparse non-negative combination of the anchors. We illustrate the performance of the method on a speaker-independent phoneme recognition task and a voice conversion task. Using a linear classifier, SABR weights achieve significantly higher phoneme recognition rates than Mel frequency Cepstral coefficients. SABR weights can also be used directly to perform accent conversion without the need to train a speaker-to-speaker regression model.

Index Terms: speech analysis, voice conversion, speaker independent representation, auditory phonetics, sparse coding

1. Introduction

Across multiple speech problems, there is a need to separate linguistic information from speaker dependent cues in the speak signal. As an example, in automatic speech recognition (ASR), speaker variability is viewed as unwanted noise in the signal of interest (i.e. linguistic content), whereas in voice conversion one seeks to modify speaker-dependent cues while retaining the linguistic content of the utterances. Unfortunately, separating these sources of information in the speech signal is a challenging task, mainly due to their complex interaction in the spectral domain [1].

Several techniques have been developed to remove physiological influences in speech. In the classical source-filter model [2] the speech signal is decomposed into source excitation, which captures the speaker's glottal characteristics, and spectral envelope. Though the spectral envelope captures (primarily) the phonetic content of the utterance, it also contains speaker-dependent information (e.g. vocal tract length). If speech recognition is the goal, vocal tract length normalization [3, 4] and speaker adaptation [5, 6] can be very effective in removing speaker dependencies from the spectral envelope, but these techniques cannot be used for source separation.

This paper presents SABR (Sparse, Anchor-Based Representation), an analysis technique that decomposes the speech signal into a set of speaker-dependent acoustic anchors and a complementary set of speaker-independent interpolation weights. Specifically, SABR uses Lasso regression [7] to approximate each acoustic frame as a sparse, non-negative linear combination of acoustic anchors. As we will show, by selecting the phoneme centroids of each speaker as anchors the resulting weights become speaker-independent. We illustrate the ability of the model to separate speaker and linguistic information on two independent problems. First, we show that SABR weights outperform conventional spectral features (MFCCs) on a speaker-independent phoneme discrimination problem. Second, we show that, by combining SABR weights derived from a source speaker with acoustic anchors from a target speaker, our technique can be used as a low-cost voice conversion method–one that does not require training a specific model for each source-target pair.

The rest of the paper is organized as follows. Section 2 reviews recent work on speech representations and their applications. Section 3 presents the SABR model and how to use its components for voice conversion and speech recognition applications. Section 4 provides details on the corpus and acoustic features used to evaluate the model, whereas section 5 presents experimental results on phonetic classification and voice conversion (subjective and objective comparison). The article concludes by discussing the implications of the results, future improvements to the method and its potential application to other speech areas.

2. Literature review

In speech recognition, a recent approach to remove unwanted speaker-specific variations is to map acoustics into the articulatory feature space. As an example, Frankel et al. [8] trained multi-layer perceptrons to estimate phonological articulatory features (e.g. place, manner, nasality, etc.) from the PLP cepstum. When they combined the estimated articulatory features with acoustic features, word error rate dropped from 67.7% to 59.7% in a speaker-independent phoneme classification task. Similarly, Arora and Livescu [9] used canonical correlation analysis (CCA) of simultaneous acoustic and articulatory recordings to capture the common factor (i.e. linguistic content) in these two views. The authors learned CCA transforms from a group of speakers and used them to extract linguistic features from acoustics in a speaker-independent fashion. CCA features improved the accuracy by 10-23% in a speaker-independent phoneme recognition task.

Articulatory features have also been used as speaker-independent representations for speech synthesis and voice conversion, but this involves building a speaker-specific mapping from articulatory features to acoustics. For example, Bollepali et al. [10] developed a speaker-specific encoder (i.e., articulatory inversion) to map acoustic features to phonological articulatory features and a decoder (i.e., forward mapping) for reverse mapping, then used the source encoder to estimate articulatory features from source utterances and decoded back to the target's acoustics using the target's decoder. Subjective tests indicated the method was successful in matching the target speaker's voice identity. In contrast, the proposed anchor-based representation in this paper obviates having to train such speaker-specific mappings.

Although not common in speech analysis, representing acoustic features as an interpolation between anchor points has been used in a handful of studies. Dijk-Kappers and Marcus [11] proposed a method of representing speech in terms of acoustic target vectors and temporal target functions. When the target vectors were derived from phonetic realizations, the authors found that phonemes could be represented using combinations of just one or two target vectors. Sun [12] characterized a sequence of acoustic feature vectors by interpolating among a set of anchor points using a smoothing spline. The anchor-based method resulted in phoneme recognition accuracy comparable to that of context-dependent HMMs, while requiring a fraction of model parameters and computational costs. In a study on music content analysis, Klapurei et al. [13] represented musical sounds by interpolating between spectra taken from appropriate temporal positions of an input signal. The method offered a better SNR for a compact representation of music than a baseline quantization method. However, the study did not explore using the method to capture speaker identity.

3. Methods

3.1. Anchor-based representation

Our proposed method represents the speech signal as a collection of speaker-dependent acoustic anchors (derived from phonetic labels) and a matrix of interpolation weights, one set of weights per acoustic frame. In this fashion, as the weights capture the similarity of each acoustic frame to various phonetic anchors, they also capture the linguistic content of the utterance, including the effects of coarticulation. Formally, SABR represents utterance X as:

$$X = A_S W \qquad (1)$$

where each column in matrix X represents an analysis window (i.e., a vector of MFCCs), A_S is a matrix of anchors for speaker S, and W is the utterance's weight matrix. If there are M acoustic frames in an utterance, N acoustic features, and P speaker anchors, then $X \in \mathbb{R}^{N\times M}$, $A_S \in \mathbb{R}^{N\times P}$, and $W \in \mathbb{R}^{P\times M}$.

3.2. Anchor selection

Several methods may be used to select the acoustic anchors in A_S, including unsupervised (e.g., k-means clustering) and supervised learning (e.g., orthogonal least-squares [14]). However, for the weight matrix W to be speaker-independent the acoustic anchors must be consistent across speakers. For this reason, SABR uses the acoustic centroid for each phoneme in the speaker's corpus as anchors –one anchor per phoneme. In this way, we argue, the sparse weights capture the linguistic content of the utterance (i.e., which phones were produced, when and how) whereas the acoustic anchors capture the identity of the speaker (i.e., voice quality and dialect/accent). As we will see, using phoneme centroids as anchors also makes the decomposition interpretable.

3.3. Sparse representation

Given a set of acoustic anchors A_S, obtained from a phonetically transcribed corpus for the speaker, and a new utterance X, we seek to find a set of weights that minimize the reconstruction error $||X - A_S W||$. A straightforward approach is to use the least-squares solution:

$$A_S^+ X = W \qquad (2)$$

where $A_S^+$is the pseudoinverse of A_S. This solution, however, does not exploit the sparse nature of the speech signal, according to which only a few anchors in A_S may be required to accurately reconstruct a given acoustic frame. Moreover, the pseudo-inverse solution allows the weight vector to take negative values, which affects the interpretability of the solution.

For these reasons, SABR enforces a sparse non-negative constraint on the solution by using Lasso regression [7]:

$$\min_\alpha \left|\left|X - A_S W\right|\right|^2 + \lambda \left|\left|W\right|\right|_1 \; s.t. W \geq 0 \qquad (3)$$

where λ is a parameter that penalizes solutions with large L1 norm. Combined with the constraint that all entries in W be nonnegative, the λ penalty term promotes sparsity (i.e., most of the entries in W are zero). For this paper, we use the Lasso implementation in the Spasm sparse coding toolbox [15].

3.4. Voice conversion with SABR

SABR provides a simple means of performing voice conversion. Given an utterance X_S from a source speaker, we first derive a set of interpolation weights (W_S) relative to the source speaker's anchors (A_S) via eq. (3). Then, given a target speaker with acoustic anchors A_T, the target speaker's utterance X_T can be estimated as:

$$X_T = A_T W_S \qquad (4)$$

As weights W_S contain phonetic information, the resulting spectrum is an estimation of the utterance said by the source speaker, but with the target speaker's voice quality.

4. Corpus

We evaluated SABR on speech from the ARCTIC speech corpus [16] which includes phonetic transcriptions for each utterance. To reduce the effect of pronunciation and phonetic differences, we chose the four native English speakers in ARCTIC as the basis for our comparison: BDL (male), CLB (female), RMS (male), and SLT (female). For each speaker, we used utterances in the "A" set to compute the SABR anchors, and utterances in the "B" set for testing purposes.

For each utterance, we used STRAIGHT [17] to extract aperiodicity, fundamental frequency and spectral envelope, then computed 24 MFCCs (24 filterbanks, 8 KHz cutoff, 15ms window, 1ms shift) from the STRAIGHT spectral envelope. We assigned each frame a phonetic label based on the ARCTIC transcription, then used $MFCC_{2-24}$ and their deltas as acoustic features, ignoring $MFCC_1$ as it contains the speech energy.

5. Experiments

Using the corpus and features discussed in the previous section, we evaluated the average Mel Cepstral Distortion[1] (MCD) between 100 target utterances and their respective voice-conversions for each combination of source and target

[1] Since voice conversions follow the timing of the source speaker, they are time-aligned to the target utterance (via dynamic time warping) prior to computing the MCD.

speakers (12 pairs). As a baseline, we also calculated the within-speaker reconstruction error. Results are shown in Figure 1. As expected, MCDs are lower when reconstructions are within-speaker than between-speakers. Additionally, the MCD is minimized at $\lambda = 0$ in the within-speaker case, which indicates that sparsity offers no benefits in this case. In contrast, the MCD in the cross-speaker case (i.e., voice conversion) is minimized at $\lambda = 0.025$, which suggests that sparsity does improve generalization across speakers. For this reason, the remaining analyses in the paper were conducted using the sparsity penalty $\lambda = 0.025$.

5.1. Phoneme classification

In a first set of experiments, we evaluated the extent to which SABR captures phonetic information in a speaker-independent manner. For this purpose, we compared SABR weights against conventional MFCC features on a phone recognition problem. Namely, we built four phoneme classifiers for each of the four ARCTIC speakers:

- **MFCC-W**: within-speaker phoneme classifier on MFCC features ($MFCC_{2-24} + \Delta_{2-24}$)
- **SABR-W**: within-speaker classifier on SABR weights (40 weights: ARCTIC phone set, excluding pause and silence frames)
- **MFCC-X**: cross-speaker classifier on MFCC features, trained on three speakers and tested on the fourth speaker
- **SABR-X**: cross-speaker classifier on SABR weights, also trained on three speakers and tested on the fourth speaker

Within-speaker classifiers were trained using 500 utterances from each speaker's training set and evaluated on test utterances from that same speaker using 8-fold cross-validation. In turn, cross-speaker classifiers were trained on the same 500 utterances from each of three speakers and tested on utterances from the excluded fourth speaker. Results are shown in Figure 2. Classification performance for the MFCCs degrades significantly when comparing within-speaker (43%) and between-speaker (23.9%), whereas classification performance for SABR features remains relatively stable: 36% versus 34.6%. Moreover, whereas MFCC features outperform SABR features by a large margin (43% versus 36.1%) in the case of within-speaker phoneme recognition, in the between-speaker case SABR features outperform MFCC features by a larger margin (34.6% versus 23.9%). These results suggest that SABR features are relatively speaker-independent. Results on the voice conversion task (discussed next) corroborate this conclusion.

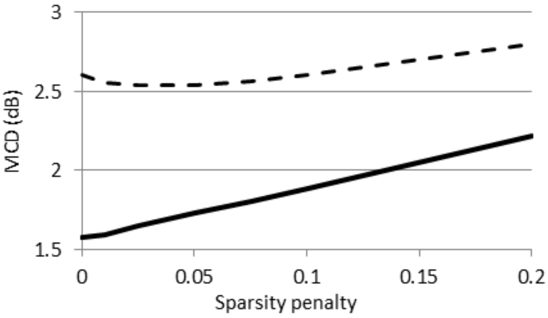

Figure 1: *Reconstruction error (MCD) within (solid line) and across speakers (dashed line). A minimum MCD exists at* $\lambda = 0.025$ *in the case of cross-speaker reconstruction (i.e., voice conversion). We computed the MCD ignoring the first coefficient, as in* [18].

5.2. Voice conversion performance

In a second set of experiments, we evaluated the ability of SABR to separate voice-quality and phonetic information using objective and subjective measures on a voice conversion task. For a particular source-target speaker pair, we used eq. (4) to reconstruct the STRAIGHT spectral envelope of the target speaker, combined it with the source energy ($MFCC_1$) and source pitch contour (scaled to match the range of the target speaker), and resynthesized the utterance with STRAIGHT.

5.2.1. Objective evaluation

First, we compared SABR against a baseline voice conversion system based on Gaussian mixture models (GMM) [19]. To control for model complexity, we limited the GMM to 40 mixtures—the number of SABR anchors. Prior to building the voice conversion model, we selected 200 training utterances using a greedy forward-selection method that maximized the entropy of the phonetic transcriptions of the utterances. Using these 200 utterances, we then build pairwise GMMs for each pair of source and target speakers (12 pairs of speakers) and computed SABR anchors for the four speakers. Results are shown in Table 1; using the 200 carefully-selected training sentences, the GMM method outperformed the SABR method on test utterances (an average MCD of 2.26 versus 2.53, respectively), likely due to the fact that each GMM was optimized for each pair of speakers and had additional free parameters (e.g. full diagonal matrices).

For this reason, we also compared the two voice-conversion models with decreasing corpus size: 100, 50, 25, and 20 training utterances selected from the corpus using the same greedy forward-selection strategy. Results are also shown in Table 1: whereas the GMM performance decreases as the number of training utterances is reduced, the SABR performance remains relatively stable. This is a promising result, particularly when considering that SABR does not require any model training, only knowledge of the source and target anchors.

5.2.2. Subjective experiments

In a final experiment, we conducted a listening test to compare the voice similarity between the SABR voice conversions and

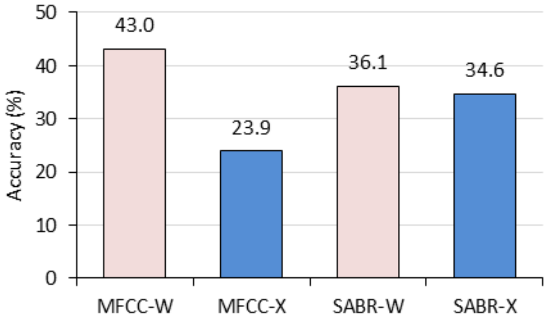

Figure 2: *phoneme classification. Performance for MFCC features degrades significantly from within-speaker to cross-speaker tasks, whereas SABR features remain stable and outperform MFCCs in the cross-speaker task.*

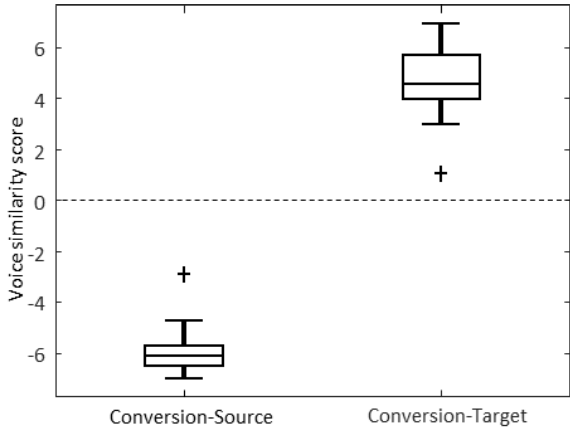

Figure 3: *voice similarity assessment results. The plot is shown on a 7-point Likert scale, rating voice similarity.*

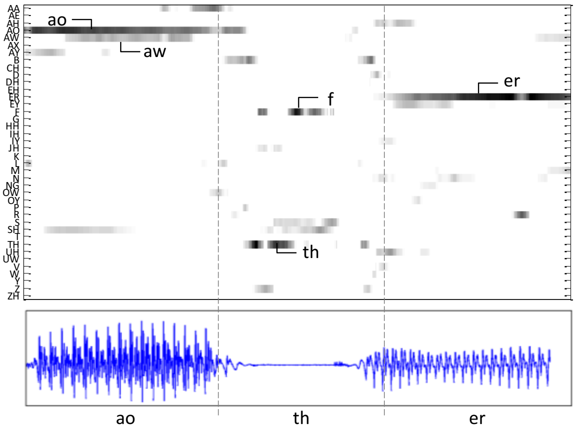

Figure 4: *example weight matrix with phonetic transcription. The weights capture the word "author" and are interpretable, particularly for vowels where the weights closely match the transcription.*

the respective source and target speakers. To account for the loss of quality due to the sparse nature of SABR synthesis, we resynthesized *source* and *target* utterances using the speaker's own phonetic anchors.

Participants were presented with 48 pairs (*source-VC* and *VC-target*) for all 12 possible speaker combinations, randomly ordered, then were asked to (1) determine if the utterances were from the same or a different speaker, and (2) rate how confident they were in their assessment using a seven-point Likert scale (1: not confident at all, 3: somewhat confident, 5: quite a bit confident, and 7: extremely confident). Following prior work [20], participants' responses and confidence ratings were then combined to form a voice similarity score (VSS) ranging from -7 (extremely confident they were from different speaker) to +7 (extremely confident they were from the same speaker).

The results of this subjective test are shown in Figure 3. Participants were "quite" confident that the converted utterances had the same voice as the target speaker ($VSS = 4.6, s.e. = 0.4$) and had a different voice from the source speaker ($VSS = -5.9, s.e. = 0.3$). This suggests that the phonetic anchors in SABR analysis successfully capture the speaker's voice identity.

6. Conclusion and future work

We have presented SABR, an analysis technique that can be used to separate voice quality and linguistic contributions to the speech signal. SABR uses sparse regularization to represent speech frames as a linear non-negative combination of acoustic anchors. By using speaker-dependent phoneme centroids as anchors, the resulting weights generalize well across speakers. In particular, our results show that SABR weights yield similar phoneme recognition performance in within-speaker and between-speaker conditions, and that they outperform conventional MFCCs in the cross-speaker condition.

Table 1: *Voice conversion performance for SABR and GMM. The top row shows the number of training utterances. Entries are the average MCD.*

Training	20	25	50	100	200
MFCC	2.66	2.59	2.40	2.31	2.26
SABR	2.59	2.59	2.57	2.56	2.53

SABR provides a straightforward method for voice conversion: an utterance from a source speaker can be converted into one for a target speaker by extracting SABR weights relative to the source anchors, and combining them with anchors from the desired target speaker. *More importantly, voice conversions can be performed without having to train a specific model for each pair of source and target speakers*. Indeed, subjective listening tests show that SABR voice conversions have the same voice quality as the target speaker. Objective measures also show that SABR is more resilient to small training corpora than a baseline GMM voice-conversion technique.

At present, SABR operates in a frame-by-frame fashion. Though the resulting weight trajectories are generally smooth—see Figure 4—incorporating smoothness constraints into the sparse optimization process may improve the representation by exploiting the temporal nature of speech (e.g. using flexible least squares [21]). In particular, while SABR represents vowels well, additional work is needed to improve its performance on consonants. Choosing anchors based on gestures, as opposed to phonetic segments (see [11]), could improve the decomposition produced by SABR. Further work will also explore the use of SABR for articulatory-inversion purposes, i.e. by measuring the articulatory configuration corresponding to each acoustic anchor, eq. (4) can be used to reconstruct the entire articulatory trajectory.

7. References

[1] H. Hermansky and D. Broad, "The effective second formant F2'and the vocal tract front-cavity," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 1989, pp. 480-483.

[2] G. Fant, *Acoustic theory of speech production: with calculations based on X-ray studies of Russian articulations* vol. 2: Walter de Gruyter, 1971.

[3] E. Eide and H. Gish, "A parametric approach to vocal tract length normalization," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 1996, pp. 346-348.

[4] D. Sundermann and H. Ney, "VTLN-based cross-language voice conversion," in *IEEE Int. Symp. on Signal Processing and Information Technology*, 2003, pp. 556-559.

[5] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech & Language,* vol. 9, pp. 171-185, 1995.

[6] S. Nakamura and K. Shikano, "Speaker adaptation applied to HMM and neural networks," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 1989, pp. 89-92.

[7] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological),* pp. 267-288, 1996.

[8] J. Frankel, M. Magimai-Doss, S. King, K. Livescu, and O. Çetin, "Articulatory feature classifiers trained on 2000 hours of telephone speech," in *INTERSPEECH*, 2007, pp. 2485-2488.

[9] R. Arora and K. Livescu, "Multi-view CCA-based acoustic features for phonetic recognition across speakers and domains," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 7135-7139.

[10] B. Bollepalli, A. W. Black, and K. Prahallad, "Modelling a Noisy-channel for Voice Conversion Using Articulatory Features," in *INTERSPEECH*, 2012, pp. 2202-2205.

[11] A. M. Van Dijk-Kappers and S. M. Marcus, "Temporal decomposition of speech," *Speech Communication,* vol. 8, pp. 125-135, 1989.

[12] D. X. Sun, "Statistical modeling of co-articulation in continuous speech based on data driven interpolation," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 1997, pp. 1751-1754.

[13] A. Klapuri, T. Virtanen, and M. Helén, "Modeling musical sounds with an interpolating state model," in *European Signal Processing Conference (EUSIPCO)*, 2005, pp. 1495-1498.

[14] S. Chen, C. F. Cowan, and P. M. Grant, "Orthogonal least squares learning algorithm for radial basis function networks," *Neural Networks, IEEE Transactions on,* vol. 2, pp. 302-309, 1991.

[15] K. Sjöstrand, L. H. Clemmensen, R. Larsen, and B. Ersbøll. (2012, 2015, February 28). *Spasm: A matlab toolbox for sparse statistical modeling* [Online]. Available: http://www2.imm.dtu.dk/projects/spasm/

[16] J. Kominek and A. W. Black, "The CMU Arctic speech databases," in *Fifth ISCA Workshop on Speech Synthesis*, 2004.

[17] H. Kawahara, "STRAIGHT, exploitation of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds," *Acoustical science and technology,* vol. 27, pp. 349-353, 2006.

[18] J. Kominek, T. Schultz, and A. W. Black, "Synthesizer voice quality of new languages calibrated with mean mel cepstral distortion," in *SLTU*, 2008, pp. 63-68.

[19] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," *Speech and Audio Processing, IEEE Transactions on,* vol. 6, pp. 131-142, 1998.

[20] D. Felps and R. Gutierrez-Osuna, "Developing objective measures of foreign-accent conversion," *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 18, pp. 1030-1040, 2010.

[21] R. Kalaba and L. Tesfatsion, "Time-varying linear regression via flexible least squares," *Computers & Mathematics with Applications,* vol. 17, pp. 1215-1245, 1989.