Blind Normalization of Speech from Different Channels

David N. Levin

Department of Radiology, University of Chicago, Chicago, IL
d-levin@uchicago.edu

Abstract

We show how to construct a channel-independent representation of speech that has propagated through a noisy reverberant channel. The method achieved greater channel-independence than cepstral mean normalization (CMN), and it was comparable to the combination of CMN and spectral subtraction (SS), despite the fact that no measurements of channel noise or reverberations were required (unlike SS).

1. Introduction

1.1. The problem. Although automatic speech recognition (ASR) technology has made steady progress in recent years, existing systems with large vocabularies are sensitive to the nature of the acoustic environment. Extensive retraining is often required if the acoustic channel is altered because the noise level changes, the speaker's room or position changes, or the signal conduit changes (e.g., telephone vs room speech). This report presents a novel method of blindly removing such channel-dependence.

1.2. Conventional methods of achieving channel-independence [1]. In most commonly-encountered situations, the acoustic environment can be characterized by a convolutive impulse response function and additive noise. In the absence of noise, a sufficiently short impulse response function has the effect of a translation in cepstral space, and CMN can be used to "subtract it out". Reverberations can also be combated by correcting for the impulse response after it has been measured by playing white noise, sine waves, or a chirp through the channel.

If the noise is stationary and is not correlated with the signal, it adds a nearly constant term to the filterbank outputs. In that case, it can be removed by SS after it has been measured. However, such a measurement requires accurate discrimination between speech and no speech, which may require the help of the recognizer in the system's "back end".

The system's "back end" can also be modified to incorporate the expected effects of a channel, although this can be computationally expensive. For example, a clean speech model can be adapted to the channel of interest by maximum likelihood linear regression (MLLR) or by parallel combination of clean speech and noise models.

1.3. The proposed method of channel normalization. Unlike existing ASR systems, humans perceive the information content of ordinary speech to be remarkably invariant in the presence of channel-dependent signal transformations. Yet there is no evidence that the speaker and listener exchange calibration data or that they measure the channel's impulse response and noise. Evidently, the speech signal is redundant in the sense that listeners blindly extract the same content from multiple acoustic signals that are transformed versions of one another. In earlier reports [2-5], the author showed how to design sensory devices that have this ability to recognize the underlying similarity of time-dependent signals differing by unknown transformations (linear or non-linear). In such devices, the signal is rescaled by a non-linear function, with the form of this scale function being determined by previously encountered signal levels. The rescaled form of a signal time series is an invariant property of it in the following sense: it is unaffected if the time series is transformed by any time-independent invertible (one to one) function. In other words, the original time series and the transformed versions of it have the same rescaled form. This is because a transformation's effect on the signal level at any time is compensated by its effect on the scale function. In earlier publications, this method was illustrated by applying it to analytic examples, simulated signals, acoustic waveforms of human speech, time-dependent spectra of bird songs, and time-dependent spectra of synthetic speech-like sounds [2-3]. This report shows how the technique can be used to represent speech cepstra in a channel-independent manner (Fig. 1). This procedure does not require the explicit measurement of the characteristics of either channel (e.g., impulse response function or noise level). One only needs to have: 1) samples of a speaker's utterances from the two channels (possibly *different* utterances from each channel); 2) a few brief reference signals from each channel, which represent the same input sounds and are used to define the origin and orientation of each channel's scale function.

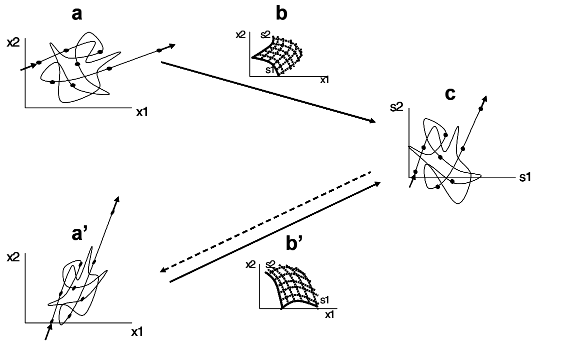

Figure 1: Schematic outline of the new method. a) The cepstral trajectory of an utterance from channel #1. b) The scale function derived from a speech sample from channel #1. a') The cepstral trajectory of the channel #2 version of the utterance in *a*. b') The scale function derived from a channel #2 speech sample. c) The trajectory found by using *b* to rescale *a*, which is also equal to the trajectory found by using *b'* to rescale *a'*. The dotted arrow shows how the channel #1 cepstra (*a*) can be converted into the channel #2 cepstra (*a'*) by mapping the rescaled values of *a* through the inverse of the channel #2 scale function (*b'*).

2. Theoretical framework

First, we argue that a time-independent invertible transformation must relate the pair of cepstral time series produced by the same utterance propagating through two time-independent channels. We make use of the embedding theorem that is well known in the field of non-linear dynamics [6]. This theorem states that almost every mapping from a *d*-dimensional space into a space of more than *2d* dimensions is invertible. Essentially, this is because so much "room" is provided by the "extra" dimensions of the higher dimensional space that the *d*-dimensional subspace, which is the range of the mapping, is very unlikely to self-intersect. Now, consider a speech signal that forms the input of any channel with stationary impulse response and noise. Because speech has 3-5 degrees of freedom [7], the power spectra of this input signal lie in a 3-5-dimensional subspace within the space of all possible power spectra. For the linear channels described in Section 1.2, the cepstral coefficients of the channel's output signal are time-independent functions of the input power spectra, and they lie in a 3-5-dimensional subspace within the space of all possible cepstra. The embedding theorem implies that there is an invertible mapping between input power spectra and the channel's output cepstra, as long as we are using a sufficient number (more than 6-10) of cepstral coefficients. Therefore, if the same input signal propagates through two different channels, the pair of output cepstral time series will be related by an invertible mapping, because each of them is invertibly related to the same time series of input power spectra. As is well known [1], this transformation between cepstra is quite non-linear if noise is present and/or if the channel's transfer function varies significantly across individual filterbank elements.

Let $x(t)$ ($x_k, k = 1,2,...,N$) be the time-dependent function that describes the trajectory of *N* cepstral coefficients of speech from a channel. In the following, we show how a special coordinate system (or scale) $s(x)$ is determined by a differential geometry that the speech trajectory imposes on the *x* manifold. Speech is invariantly represented in this coordinate system in the following sense: if its cepstral trajectory is subjected to any invertible transformation, the representation of the transformed trajectory in *its s* coordinate system is the same as the representation of the untransformed speech in *its s* coordinate system. To see how this comes about, consider a point *y* in a region of the *x* manifold that is densely sampled by the speech trajectory. Define g^{kl} to be the average outer product of the time derivatives of the speech trajectory as it passes through a small neighborhood of *y*: $g^{kl} = \left\langle \frac{dx_k}{dt}\frac{dx_l}{dt} \right\rangle_{x(t)\sim y}$, where the bracket denotes the average over time. As long as this neighborhood contains *N* linearly independent time derivatives, g^{kl} is positive definite, and its inverse g_{kl} is well defined and positive definite. Under any change of coordinate systems $x \rightarrow x' = x'(x)$, $\frac{dx}{dt}$ transforms as a contravariant vector. Therefore, g^{kl} and g_{kl} transform as a contravariant and covariant tensors, respectively. This means that g_{kl} can be taken to define a metric on the *x* manifold, and a coordinate-independent process for moving (parallel transporting) vectors across the manifold can be derived from this metric by means of the methods of Riemannian geometry. Now suppose that *N* linearly-independent "reference" vectors h_a $(a = 1,2,...,N)$ can be defined at a special "reference" point x_0 on the manifold. For example, in the experiments in Section 3, each reference vector was taken to be the average cepstral velocity during one of a few brief manually-chosen time intervals when the speech trajectory passed through the same neighborhood in cepstral space. The reference vectors can be parallel transported across the manifold to define the *s* coordinates of any point *x*. For example, the point *x* can be assigned the coordinates s ($s_k, k = 1,2,...,N$) if it is reached by starting at x_0 and then: parallel transporting h_1 along itself s_1 times while simultaneously parallel transporting the other h_a along the same path, then parallel transporting h_2 along itself s_2 times while simultaneously parallel transporting the other h_a along the same path, …, and finally parallel transporting h_N along itself s_N times. Notice that this parallel transport process is independent of what coordinate system is used on the cepstral (*x*) manifold. Therefore, as long as the reference point/vectors can be identified in a coordinate-independent manner, the *s* representation of the speech trajectory will also be coordinate-independent. Because an invertible transformation of the trajectory is mathematically equivalent to a change of the manifold's coordinate system, this means that speech trajectories related by invertible transformations will have the same *s* representation. Recall that the embedding theorem implies that there is an invertible mapping between the speech trajectories of an utterance propagating through two different channels. It follows that these trajectories have identical *s*-representations. This representation can be used directly as channel-independent input of a recognizer. Alternatively, as shown in Fig. 1, this procedure can be used to perform channel conversion: i.e., to modify the cepstral time series of speech from one channel (a corrupted channel) so that it resembles the cepstral time series of the same utterance from another (clean) channel, on which the system was trained.

In the above discussion, it was assumed that the two scale functions were derived from identical utterances that had propagated through the two channels. However, suppose it is assumed that different utterances from the same speaker/channel combination always lead to the same metric and scale function. Then, the above channel conversion procedure can be performed even if different speech samples have been observed in the two channels. In other words, one can use the scale functions derived from different clean and corrupted utterances to predict the cepstral coefficients of the clean versions of corrupted utterances. The success of the experiments in Section 3 suggests that speech scale functions have this property of utterance-independence; i.e., they are stable with respect to speech content. This is not surprising for the following reason. We know that speech is composed of a small number of units (e.g., phonemes) that occur repeatedly with certain frequencies. Therefore, two sufficiently large samples of speech are likely to produce the same distribution of cepstral velocities in each cepstral neighborhood. Because the metric reflects the statistical distribution (i.e., the covariance matrix) of those velocities, the two speech samples will also lead to the same metric and the same scale function.

3. Experimental results

We performed experiments on data from three male and female speakers of American English, who were part of the DARPA Air Travel Information System (ATIS0) corpus of speaker-dependent training data [8] and who represented different accent regions. The ATIS0 speech samples were recorded with a Sennheiser microphone at a 16 kHz sampling rate with 16 bits of depth. For each speaker, the clean speech sample was comprised of the unmodified data representing 12 sentences (approximately 80 s) of this corpus. Non-overlapping sets of sentences were used to define the clean speech samples of different speakers. The acoustic waveform of each sentence was Fourier transformed, after it had been Hamming-windowed in 24 ms time frames at 4 ms intervals. Each frame's power spectrum was used to compute 20 mel frequency cepstral coefficients (MFCC). For each speaker, the set of 12 sentences defined a time series of approximately 2 x 10^4 cepstra, which formed a trajectory in cepstral space. This trajectory densely traversed and retraversed a compact "speech domain", whose location, size, and shape depended on the speaker and channel characteristics. The speech trajectory was dimensionally reduced by retaining its first two principal components, which contained approximately 95% of the data's variance (Fig. 2).

The trajectory was covered with a uniform 64 x 64 array of rectangular neighborhoods within which the clean speech metric was computed by the formula in Section 2. Then, parallel transport was defined in terms of an affine connection, which was given by the standard combination of metric derivatives in the Christoffel bracket. For each speaker, we manually identified a tight cluster of cepstra that represented brief sounds in the clean speech sample (each sound being 4 ms long). These were used to determine a reference point and reference vectors (x_0 and h_a) that defined the origin and local orientation of the clean speech scale. Then, the complete scale (Fig. 2) was formed by parallel transporting these reference vectors away from the origin, as described in Section 2.

For each speaker, a corrupted speech sample was created from 12 *different* sentences by convolving each signal with a channel impulse response function and adding Gaussian white noise in the time domain. *Note that no sentence of the ATIS0 corpus was used twice for the same speaker or for different speakers.* Each speaker's speech was corrupted by one of two impulse response functions, which were synthesized by the "image source" method [9]. These functions described small rooms with different levels of reverberation (reflectivity ~ 0.7-0.9) in which the speaker and microphone were separated by varying distances (25-112 cm). Each impulse response included all reverberations with echo times less than 64 ms. After addition of noise, the SNR of the corrupted speech was 16-20 db in each case. As above, the acoustic waveform of the corrupted speech was used to compute an MFCC time series, which formed a trajectory in cepstral space (Fig 2). This data was dimensionally reduced by retaining its first two principal components (containing approximately 89% of the data's variance), and the metric and affine connection of corrupted speech were computed. Corrupted versions of the clean speech reference sounds were used to determine the corrupted reference information (x'_0 and h'_a), and the corrupted speech scale was then defined by parallel transporting these reference vectors away from the origin. It is important to note that these brief reference sounds were the only information that was common to the derivations of the clean and corrupted speech scales, which were otherwise based on entirely different sets of utterances. Notice that the scale function for corrupted speech (Fig. 2) is ill-defined in the lower half of the speech domain because of the relative paucity of data there.

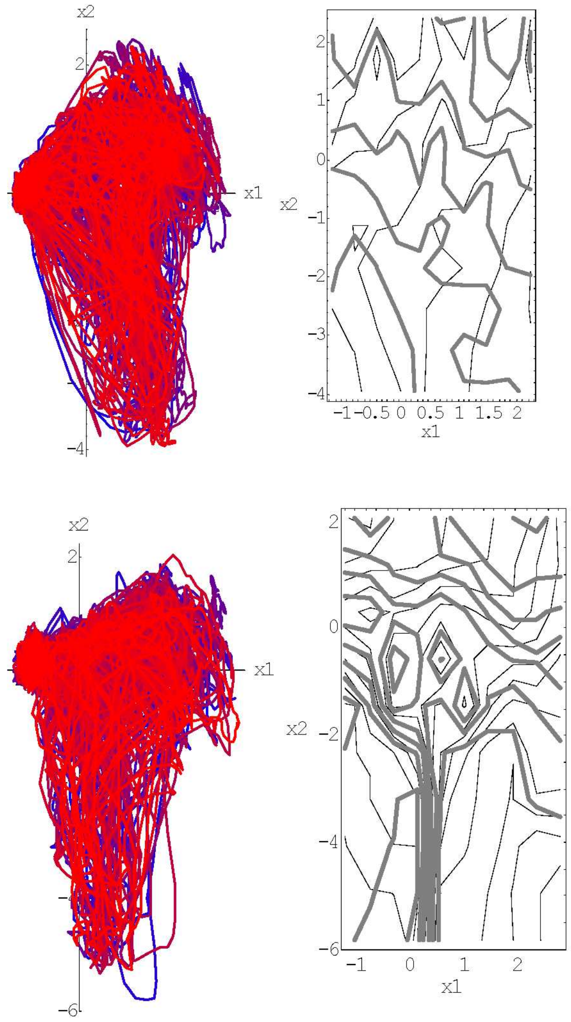

Figure 2: Left: The trajectories of the first two principal components of the cepstra of 12 clean (top) and corrupted (bottom) sentences from speaker BF. These figures have been rotated and rescaled along each axis to show detail. Right: The scale functions derived from the left panels. The thin black (thick gray) lines are s_2 (s_1) isoclines.

Next, the scales of clean and corrupted speech were used to perform the channel conversion process described in Section 2 (Fig. 1). Figure 3 shows a typical result. Notice that the channel-converted MFCCs and the clean MFCCs were much closer to one another than were the corrupted and clean MFCCs after normalization by CMN. Figure 3 also shows the distributions of Euclidean distances between the corrupted and clean MFCCs (after CMN) and between the channel-converted and clean MFCCs, at 1430 time points during all words in

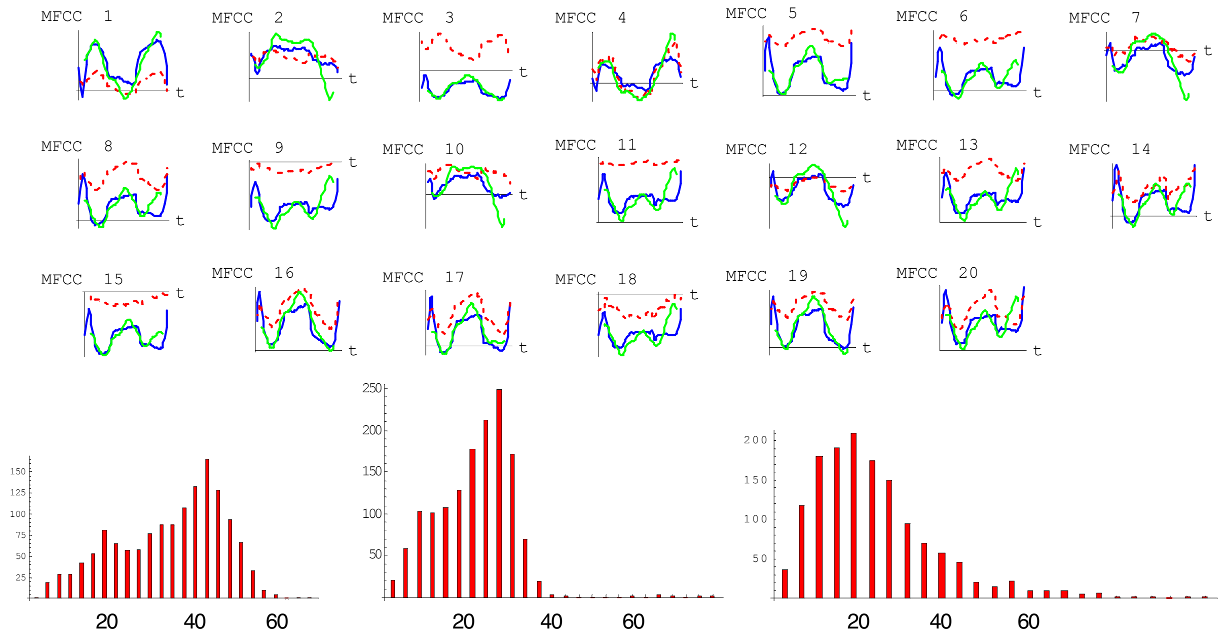

Figure 3: Upper: The solid black (blue) and dashed (red) lines show the MFCCs of the clean and corrupted versions of the words "and make", respectively, after "normalization" by CMN. Solid gray (green) lines show the corrupted MFCCs after the new channel conversion procedure. Lower: The distribution of Euclidean distances between the corrupted and clean cepstra after CMN (left), after CMN + SS (center), and after the new channel conversion process (right). The means (99% confidence intervals) of these distributions are: 35.4±0.9, 22.9±0.6, and 23.4±1.0.

three typical sentences. These histograms show that the channel conversion process did a much better job than CMN in moving the corrupted MFCCs close to the clean MFCCs at the great majority of time points. Furthermore, the new channel conversion procedure was comparable to the combination of CMN + SS in its ability to normalize speech from different channels. This is true despite the fact that the channel conversion procedure did not involve the measurement of noise levels required by SS. Very similar results were obtained for the other speakers.

A technical comment should be made at this point. Recall that the scales of clean and corrupted speech were derived from dimensionally-reduced data. Therefore, in Fig. 3, we compared the ability of the channel conversion process to the ability of conventional methods (CMN alone or CMN + SS) to predict the dimensionally-reduced clean MFCCs from dimensionally-reduced corrupted MFCCs. However, we also found that the Euclidean distances between fully-dimensional clean MFCCs and corrupted MFCCs were reduced more by the above channel conversion procedure than by CMN.

4. Discussion

Previous publications [2-5] described a novel method of representing signal time series that essentially "filters out" the effects of unknown distortions. In this report, the method was used to create channel-independent representations of speech cepstra. The experimental results suggest that the new technique is more successful than CMN and comparable to CMN + SS in its ability to decrease the signal's channel dependence. Even better results can be expected if more of the data's variance is retained in the dimensional reduction step and if longer speech samples are used to compute the metric and scale. Notice that the new method has the following advantages compared to conventional approaches to channel normalization: 1) it does not require prospective measurements of the channel's impulse response and noise; 2) it is a pure "front end" technology and avoids the computational demands of modifying or retraining the system's recognizer. In principle, an ASR system with the new front end can be trained in one environment and then used in another without additional measurements or retraining.

References

[1] Huang, X., Acero, A., and Hon, H-W., *Spoken Language Processing*, Prentice Hall, Upper Saddle River, NJ, 2001.

[2] Levin, D. N., "Sensor-Independent Stimulus Representations", *Proc. Nat'l. Acad. Sci. (USA)*, 99: 7346-735, 2002.

[3] Levin, D. N., "Representations of Sound That Are Insensitive to Spectral Filtering and Parameterization Procedures", *J. Acoust. Soc. Am.*, 111: 2257-2271, 2002.

[4] Levin, D. N., "Blind Normalization of Speech from Different Channels and Speakers", *Proc., 7th Internat Conf on Spoken Language Process.*, Denver, CO, September 16-20, 2002.

[5] Levin, D. N., "Blind Normalization of Speech from Different Channels", *Proc., COST 277 Workshop on Non-Linear Processing of Speech*, Edinburgh, Scotland, December 2-3, 2002.

[6] Sauer, T., Yorke, J. A., and Casdagli, M., "Embedology", *J. Stat. Phys.*, 65: 579-616, 1991.

[7] Tishby, N., "A Dynamical Systems Approach to Speech Processing", *Proc., 1990 Internat. Conf. on Acoustics, Speech, and Signal Process.*, 1: 365-368, 1990.

[8] Linguistic Data Consortium: http://www.ldc.upenn.edu.

[9] Allen, J. and Berkeley, D., "Image Method for Efficiently Simulating Small Room Acoustics", *J Acoust. Soc.Am.*, 65: 943-950, 1979.