BLIND NORMALIZATION OF SPEECH FROM DIFFERENT CHANNELS

David N. Levin

Dept. of Radiology, U. of Chicago and Invariant Sensor Technologies, Inc.

ABSTRACT

We show how to construct a channel-independent representation of speech that has propagated through a noisy reverberant channel. This is done by blindly rescaling the cepstral time series by a non-linear function, with the form of this scale function being determined by previously encountered cepstra from that channel. The rescaled form of the 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 distortion. Because a linear channel with stationary noise and impulse response transforms cepstra in this way, the new technique can be used to remove the channel dependence of a cepstral time series. In experiments, the method achieved greater channel-independence than cepstral mean normalization, and it was comparable to the combination of cepstral mean normalization and spectral subtraction, despite the fact that no measurements of channel noise or reverberations were required (unlike spectral subtraction).

1. INTRODUCTION

1.1 The problem.

Despite the steady progress of speech recognition technology in recent years, existing systems with large vocabularies are still sensitive to the nature of the acoustic environment and to the identity of the speaker [1]. For example, 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 (telephone vs. room speech). This report describes a non-linear signal processing method that makes speech signals more channel-independent and that can be used in the "front end" of any ASR system.

1.2. Conventional approaches to channel-independent ASR.

In most commonly-encountered situations, the acoustic environment can be characterized in the time domain by a convolutive impulse response function and additive noise. In this case, the corrupted speech signal is parameterized by the filterbank outputs:

$$P_i = \int |X(f)|^2 |H(f)|^2 M_i(f) df + N_i \qquad (1)$$

where P_i is the power of the corrupted signal from the i^{th} filterbank element, $|X(f)|^2$ is the power density of the channel's input (clean) signal, $|H(f)|^2$ is the power density of the channel's impulse response function, $M_i(f)$ is the profile of the i^{th} filterbank element, and N_i is the noise power from that element. This equation depends on the following approximations, which are commonly made and often work well in practice [1]: 1) the impulse response is small at time delays greater than the length of the spectral window; 2) the noise is not correlated with the speech. Notice that the noise term in Eq.(1) represents the noise power integrated over relatively wide filterbank elements (e.g., elements of a mel frequency filterbank). Therefore, to the extent that the underlying noise distribution is stationary and "white", this term is an average quantity with small frame-to-frame fluctuations.

Now, suppose that an ASR system was trained to recognize speech in one environment (e.g., clean speech) and it is now being used to analyze utterances from another channel (e.g., corrupted speech). If the channel transfer function H is approximately constant over each filterbank element, it can be factored out of the integral in Eq.(1). Then, in the absence of noise, it simply has the effect of a translation in cepstral space, and cepstral mean normalization [2] (CMN) can be used to "subtract it out" in order to remove the effects of reverberations. However, if noise is present and/or the transfer function is narrow, there is a non-linear relationship between the cepstra from the two channels, and CMN is not as effective.

The simplest way of accounting for noise is spectral subtraction (SS), but this requires periodic noise power measurements [3]. Therefore, its implementation requires accurate discrimination between speech and no speech, which may require the help of the recognizer in the system's "back end".

Another approach is to modify the system's back end in order to incorporate the expected effects of a channel [1]. For example, in "multi-style training", the recognizer is trained on a database that contains speech samples from a variety of common channels. In principle, this method has the disadvantage of "blurring" the statistical distributions of the recognizer, and, of course, it may perform poorly in the presence of an unanticipated channel. Alternatively, a clean speech model can be adapted to the channel of interest by using maximum likelihood linear regression or by a parallel combination of clean speech and noise models. However, this may entail a significant computational expense.

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 [4-9], the author showed how to design sensory devices that have this ability to normalize time-

dependent signals differing by unknown transformations (linear or non-linear). In such devices, the signal is blindly 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) distortion. 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, spectral time series of bird songs, and spectral time series of synthetic speech-like sounds [4-6, 9].

This approach is relevant to speech recognition for the following reason. After an utterance is passed through two different channels, the cepstral time series of the resulting output signals are related to one another by a non-linear transformation that characterizes the differences between the two channels. As shown in Section 2, Eq.(1) implies the invertibility and time-independence of this transformation, as long as each channel's impulse response and noise distribution are stationary. Therefore, the same representation will result when the cepstral time series from either channel is blindly rescaled by the new signal processing method (Fig.1). Alternatively, the cepstral time series of utterances from channel #1 can be used to estimate the cepstral time series of the same utterances from channel #2 by first finding the invariant representation of the channel #1 signal and then synthesizing the channel #2 signal having the same invariant representation (dotted arrow in Fig. 1). This "channel conversion" procedure can be performed as long as one has: 1) samples of a speaker's utterances from the two channels (possibly *different* utterances from each channel); 2) a few reference cepstra from each channel, which represent the same input sounds and are used to define the origin and local orientation of each channel's scale function. Notice that the new method has the following advantages compared to conventional approaches to channel normalization: 1) it does not require explicit 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.

2. THEORY

In this section, 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. Then, we demonstrate how these two cepstral time series can be blindly rescaled so that they have the same representation. This rescaling process can be used to perform channel conversion: i.e., to modify the cepstral coefficients of an utterance from one channel so that they resemble those of the same utterance from another channel.

We make use of the embedding theorem that is well known in the field of non-linear dynamics [10]. 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

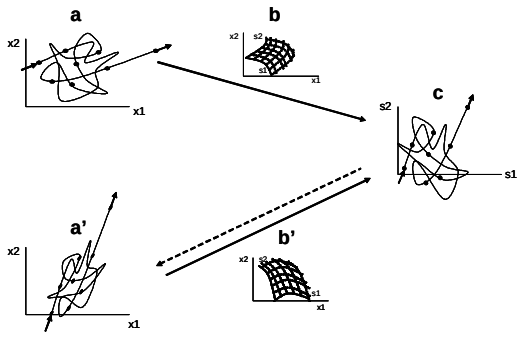

Figure 1. Schematic outline of the new method. a) The cepstral trajectory of an utterance from channel #1. b) The scale function derived from a speech sample from channel #1. a') The cepstral trajectory of the channel #2 version of the utterance in *a*. b') The scale function derived from a channel #2 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'*).

"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 is thought to have 3-5 degrees of freedom [11] (S. Parthasarathy, AT&T Labs, private communication, 2001), 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 (Eq. (1)), and they lie in a 3-5-dimensional subspace within the space of all possible cepstra. The embedding theorem implies that this mapping is invertible, as long as we are using a sufficient number of cepstral coefficients (more than 6-10). 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.

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 the speech normalization problem can be mapped onto the geometric problem of finding an invariant description of non-linearly related cepstral trajectories, and the latter problem is then solved with the help of differential geometry. Specifically, 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*: $g^{kl} = \left\langle \frac{dx_k}{dt} \frac{dx_l}{dt} \right\rangle$, where the bracket denotes the average over times when $x(t)$ is close to *y*. 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, $x \to x' = x'(x)$, $\frac{dx}{dt}$ 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. For instance, the parallel transport process can be defined by means of an affine connection equal to the Christoffel symbol, which is composed of products of the metric's derivative and the inverse metric [12]. 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, we identified all trajectory segments connecting consecutive cepstral points in a small neighborhood of cepstral space, found *N* linear combinations of their cepstral velocities that were non-vanishing and linearly-independent, and defined those averages to be the reference vectors. The reference vectors can be parallel transported across the manifold to determine the *s* coordinates of any point *x*. Specifically, the point *x* can be assigned the coordinates *s* ($s_k, k = 1,2,...,N$), if it is reached by starting at x_0, 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-system-independent manner, the *s* representation of the speech trajectory will also be independent of the choice of coordinate system. 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 the existence of an invertible mapping between the pair of cepstral trajectories of an utterance that propagated through two different channels. It follows that these trajectories have identical *s*-representations (Fig. 1). In principle, this representation can be used directly as channel-independent input of a recognizer.

However, in this report, we use this procedure 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 (Fig. 1). Then, the converted cepstral coefficients can be fed into a conventional recognizer that has been trained on clean speech. To see how this is done, let $x(t)$ be the cepstral time series of an utterance from channel #1, and let $s(x)$ be the scale function derived from a speech sample from channel #1. Likewise, let $x'(t)$ be the cepstral trajectory of the same utterance from channel #2, and let $s'(x')$ be the scale function derived from the aforementioned speech sample after propagation through channel #2. In the previous paragraph, we showed that the rescaled representations of these two trajectories are the same: i.e., $s[x(t)] = s'[x'(t)]$. Therefore, the cepstral coefficients of the channel #2 speech can be found by mapping the *s*-representation of the channel #1 speech through the inverse of the scale function of the channel #2 speech: $x'(t) = s'^{-1}[s[x(t)]]$. Now, in the above discussion, it was assumed that the two scale functions were derived from identical speech samples that had propagated through the two channels. However, suppose 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 speech samples 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 of those velocities (i.e., the velocity covariance matrix), two speech samples will lead to the same metric and the same scale function.

3. EXPERIMENTAL RESULTS

We performed experiments on data from three speakers of American English, who were part of the Air Travel Information System (ATIS0) corpus of speaker-dependent training data [13]. As shown in Table I, these subjects were from a variety of accent regions and included males and females of various ages. 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 11 or 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 [14] (MFCC). For each speaker, the set of 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 (Fig. 2). The speech trajectory was dimensionally reduced by retaining its first two principal components, which contained approximately 95% of the data's variance.

Each 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, given by the Christoffel bracket [12]. For each speaker, we manually defined a small neighborhood with dimensions equal to 5-10% of the size of the whole speech trajectory, and we identified all clean speech trajectory segments between consecutive cepstral points in that neighborhood. The origin of the clean speech scale (x_0) was taken to be the average position of those points, and the reference vectors (h_a), which determined the orientation of the scale's axes at x_0, were derived from the cepstral velocities between those points, as in Section 2. Then, the complete scale (Fig. 2) was formed by parallel transporting these reference vectors away from the origin, as described in Section 2. Scale values in regions immediately outside the traversed speech domain were estimated by extrapolating the scale values found by parallel transport within the speech region.

For each speaker, a corrupted speech sample was created from 11 or 12 *different* sentences by convolving each ATIS0 signal with a channel impulse response function and adding stationary Gaussian white noise in the time domain. Note that different sets of clean and corrupted sentences were used for different speakers. Each speaker's speech was corrupted by one of two impulse responses, which were synthesized by the "image source" method [15]. One of these functions described a relatively "hard" or reverberant small room (reflectivity ~ 0.9), in which the speaker and microphone were "close" (25 cm apart). The other impulse response corresponded to a "softer" version of the same room (reflectivity ~ 0.7), in which the speaker and microphone were "far" (112 cm apart). Each impulse response included all reverberations with echo times less than 64 ms. After addition of noise, the SNR of the corrupted speech was 10-20 dB in each case. As above, the acoustic waveform of the corrupted speech was used to compute an MFCC time series, which was dimensionally reduced by retaining its first two principal components (containing approximately 90% of the data's variance), and the metric and affine connection of corrupted speech were computed. Corrupted versions of the clean speech reference cepstra 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 (Fig. 2). It is important to note that these reference cepstra 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 sentences. Notice that the corrupted speech scale is compressed centrally relative to the clean speech scale (Fig. 2), reflecting a similar relative compression of the corrupted speech trajectory.

Next, the scales of clean and corrupted speech were used to perform the channel conversion process described in

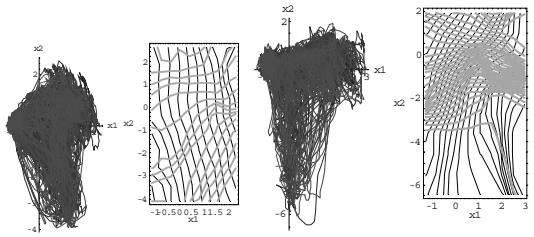

Figure 2. Panels from left to right: 1) the trajectory of the first two principal components of the cepstra of 12 clean sentences from speaker BF. This figure has been rotated and rescaled along each axis to show detail. 2) The scale function derived from the left panel. The thin black (thick gray) lines are s_2 (s_1) isoclines. 3) The trajectory of the 12 corrupted sentences from speaker BF ("soft" room, "far" microphone, SNR=10 dB). 4) The scale function derived from the third panel.

Section 2 (Fig. 1). Specifically, the MFCCs of corrupted sentences were used to predict the MFCCs of clean versions of those sentences. First, the corrupted MFCCs were rescaled with the scale function of corrupted speech. The rescaled values were then mapped through the inverse scale function of clean speech to predict the MFCCs of the clean versions of the corrupted utterances. These were compared to the MFCCs of the actual clean versions of those utterances (i.e., the original ATIS0 versions before corruption by the channel's impulse response and noise). Figure 3 shows an example of this type of comparison for the words "one way", spoken by speaker BF. It is apparent that the channel-converted corrupted MFCCs and the clean MFCCs were much closer to one another than were the corrupted and clean MFCCs after normalization by CMN. Note that this result was produced by a procedure that does not involve the variation of any free parameters in order to best fit the data. Table I lists the mean Euclidean distances between the corrupted and clean MFCCs (after CMN) and between the channel-converted corrupted and clean MFCCs during all words in three typical sentences. It is evident that the channel conversion process did a much better job than CMN in moving the corrupted MFCCs close to the clean MFCCs. 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 spectral subtraction. Similar results were obtained for the other speakers and channels (Table I).

Two technical comments should be made at this point. First, recall that the scales of clean and corrupted speech were derived from dimensionally-reduced data. Therefore, the channel conversion process is only expected to predict the *dimensionally-reduced* MFCCs of clean versions of corrupted speech. It is NOT capable of predicting higher principal components of these MFCCs. Therefore, in Fig. 3 and Table I, we compared how well the channel conversion process and conventional normalization methods (CMN alone or CMN + SS) could predict the dimensionally-reduced clean MFCCs from dimensionally-reduced corrupted MFCCs.

Another technical issue concerns the ranges of the scale functions derived from the clean and corrupted speech samples. Each of these scale functions sweeps out a range of

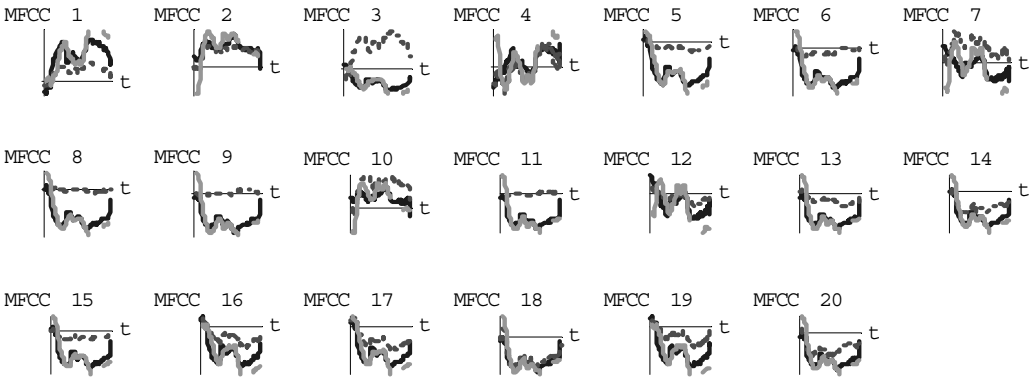

Figure 3. Speaker BF ("soft" room, "far" microphone, SNR=10 dB). The blue (dark solid) and red (dashed) lines show the MFCCs of the clean and corrupted versions of the words "one way", respectively, after "normalization" by CMN. The green (solid gray) lines show the corrupted MFCCs after channel conversion.

SPEAKER	AGE/GENDER	ACCENT	ROOM/MIC	SNR	CMN	CMN + SS	CHANNEL CONVERSION
BF	20/M	western	hard/close	16 dB	35.4±0.9	22.9±0.6	23.4±1.0
			soft/far	10 dB	35.5±1.1	26.9±0.9	27.9±1.4
B0	40/F	north midland	soft/far	16 dB	40.1±0.8	28.4±0.7	27.8±1.3
B5	30/F	south midland	hard/close	20 dB	53.8±0.6	36.7±0.4	35.5±1.1
			soft/far	14 dB	52.8±1.2	38.9±0.9	34.7±1.5

Table I. The mean Euclidean distance between clean cepstra and those corrupted by reverberations and noise, after "normalization" by CMN, CMN + SS, and the new channel conversion procedure. In each case, the cepstra describe 1068-2860 time points in all words in three typical sentences. The 99% confidence interval of each mean distance is listed.

rescaled cepstra (*s* values) over the domain of the unrescaled cepstra (*x* values) of the corresponding speech sample. In principle, these two ranges should be the same, but they differed somewhat in actual practice. Because of this, some cepstra near the edges of the corrupted speech domain (typically 35% of the total) were rescaled to values outside the range of the clean speech scale function. Therefore, they fell outside the domain of the inverse of the clean speech scale function, and they could not be mapped through that inverse in order to compute their channel-converted values (Fig. 1).

4. CONCLUSIONS

Previous publications [4-9] described a new method of representing signal time series that essentially "filters out" the effects of unknown distortions. In this paper, the method was used to blindly create relatively 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. 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 (D. N. Levin, patents pending). Of course, this hypothesis must be tested by comparing the word error rates of ASR systems with and without the new front end.

Note that this method is capable of performing adaptive channel normalization. Specifically, the system can adapt to changing channel conditions by using the most recent sample of corrupted speech in order to periodically update the metric and scale function. The updated scale of corrupted speech, together with the static scale of the clean speech, can be used to estimate the clean speech cepstra corresponding to observed corrupted speech. This adaptive rescaling technique was demonstrated successfully on one-dimensional signals in reference 6.

It should be pointed out that the ideas in this paper can be applied in more general circumstances. In Section 2, the embedding theorem was used to argue that the power spectrum of an acoustic channel's input and the MFCCs of its output are related by an invertible transformation, which characterizes the channel. By similar reasoning, the input power spectrum is invertibly related to the values of *any* set of spectral parameters that are used to characterize the channel's output power, as long as those parameters are sufficiently numerous (more than 6-10) and as long as they average the power over many frequencies. Therefore, if a channel's output is detected with two different spectral parameter measurements (e.g., MFCCs vs. linear frequency cepstral coefficients), the pair of output time series will be invertibly related to one another, because each of them is invertibly related to the same time series of input power spectra. It follows that these output time series will rescale to the same form. This means that two ASR systems will derive the same rescaled representation of a signal, even though they used different spectral parameters to "sense" it and even though they received it through two different channels. In reference 6, this insensitivity of the rescaled signal to the choice of spectral

measurements was demonstrated in experiments on bird songs having one underlying degree of freedom. The embedding theorem further guarantees that the power spectrum of the channel's input is a one-to-one function of the 3-5 parameters that characterize the instantaneous configuration of the speaker's vocal tract. It follows that the measured spectral parameters of the channel's output are also invertibly related to the speaker's vocal tract parameters. Therefore, the rescaled form of a measured spectral parameter trajectory is the same as the rescaled form of the spatial trajectory of the articulatory apparatus. In other words, although the observed time series of speech spectral parameters may not enable one to recover an absolute description of vocal tract motion in a specific laboratory spatial coordinate system, it does enable one to recover its rescaled form. In this sense, the theoretical framework of this paper is consistent with the "motor' theory of speech perception. Because an invertible mapping is mathematically equivalent to a change of coordinate systems, the measured spectral parameters can be considered to describe the speaker's vocal tract configuration in a particular coordinate system. Therefore, if two ASR systems are "listening" to a given speaker through different channels and/or are measuring different spectral parameters, they are both recording the trajectory of the speaker's vocal tract, although they are describing it in different coordinate systems. Mathematically speaking, the "inner" properties of a geometrical figure are those that are independent of the coordinate system used to numerically describe it (or, equivalently, independent of transformations of the figure in a fixed coordinate system). Geometry seeks to find these "inner" properties and is less concerned with the figure's "outer" properties: i.e., aspects of its description that depend on the choice of coordinate system. From a geometrical perspective, the trajectory of a particular set of speech spectral parameters (e.g., MFCCs) is an "outer" property of the articulatory gesture because it depicts the gesture in a coordinate-system-dependent manner. In contrast, the rescaled form of a speech signal is an "inner" property of the vocal tract motion, describing aspects of the motion that are independent of the coordinate system used to depict it (and, therefore, independent of the nature of the spectral parameters and channel).

It is interesting to consider the possibility that rescaling can be used to create *speaker-independent* representations of speech signals. This idea is outlined here, but in other reports this process was experimentally demonstrated with synthetic speech-like sounds having a single degree of freedom [6]. Suppose there is an invertible transformation that consistently maps the instantaneous configuration of one speaker's vocal tract onto the configuration of the other speaker's vocal tract when they utter the same words. This is equivalent to the assumption that the two speakers' articulatory gestures consistently mimic each other when the same words are uttered. In this case, the MFCC trajectories of the two speakers' signals must be invertibly related to one another, because each is invertibly related to the trajectory of the originating vocal tract and the configurations of the two vocal tracts are invertibly related to each other. It follows that the MFCC trajectories of the two signals have the same rescaled form; i.e., a speaker-independent form.

The rescaling procedure may be useful for addressing the problem of *speaker identification*, because it cleanly separates speech content from the characteristics of speaker and channel. Specifically, each speaker/channel combination is associated with a non-linear scale function or *s* coordinate system that covers the patch of cepstral space traversed by the signals from that source (e.g., the warped grid of *s* isoclines in Fig. 2). The speaker/channel is characterized by the location and configuration of this scale within the space of MFCCs. The content of an utterance is described by the form of its cepstral trajectory in this special coordinate system, which defines a kind of speaker/channel-dependent "medium" on which the message is "written". Although the trajectory of a given utterance in MFCC space is non-linearly transformed by a change of speaker/channel, it has an invariant form in the special *s* coordinate system associated with each speaker/channel combination, because the *s* coordinate system is non-linearly transformed in the same way.

5. REFERENCES

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

[2] B.S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification", *J. Acoust. Soc. Am.*, vol. 55, 1304-1312, 1974.

[3] S.F. Boll, "Suppression of acoustic noise in speech using spectral subtraction", *IEEE Trans. Acoustics, Speech, and Signal Process.*, vol. 27, 113-120, 1979.

[4] D.N. Levin, "Stimulus representations that are invariant under invertible transformations of sensor data", *Proc. SPIE*, vol. 4322, 1677-1688, 2001.

[5] D.N. Levin, "Sensor-independent stimulus representations", *Proc. Nat'l. Acad. Sci. (USA)*, vol. 99, 7346-7351, 2002.

[6] D.N. Levin, "Representations of sound that are insensitive to spectral filtering and parameterization procedures", *J. Acoust. Soc. Am.*, vol. 111, 2257-2271, 2002.

[7] D.N. Levin, "Blind normalization of speech from different channels and speakers", on the CD-ROM: *ICSLP-2002 Conf. Proc., 7th Int. Conf. on Spoken Language Process., Denver, CO, September 16-20, 2002* (ISBN 1-876346-40-X).

[8] D.N. Levin, "Blind normalization of speech from different channels", on the CD-ROM: *Eurospeech '03: 8th European Conf. on Speech Comm. And Tech., Geneva, Switz., September 1-4, 2003* (ISSN 1018-40-74).

[9] D.N. Levin, papers at http://www.geocities.com/dlevin2001/

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

[11] N. Tishby, "A dynamical systems approach to speech processing", *Proc., 1990 Int. Conf. on Acoustics, Speech, and Signal Process.*, vol. 1, 365-368, 1990.

[12] E. Schrodinger, *Space-Time Structure*, Cambridge U. Press, Cambridge, UK, 1963, p. 63.

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

[14] S.B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", *IEEE Trans. Acoustics, Speech, and Signal Process.* vol. 28, 357-366, 1980.

[15] J. Allen and D. Berkeley, "Image method for efficiently simulating small room acoustics", *J. Acoust. Soc. Am.*, vol. 65, 943-950, 1979.