A STUDY ON USING THE MELLIN TRANSFORM FOR VOWEL RECOGNITION

Antonio De Sena and Davide Rocchesso
Università di Verona
Dipartimento di Informatica

ABSTRACT

In this paper we discuss the applicability of the Mellin transform for vowels recognition, focusing on spectral envelope scale distribution.
The hypothesis used is that same vowels produced same spectral envelope shapes (same curves with different compression factor), so an energy, time and scale normalization can be used to map same vowels to same distributions.

So, using fast algorithms, we study the applicability of this idea to build a realtime or quasi-realtime system capable of making vowel discrimination in a relatively straightforward way.

1. INTRODUCTION

Same vowels pronounced by different people with different gender, different age, or by the same person using a different pitch can be recognized by our auditory system. We don't know how our auditory system can do this, but we can try to find a way to replicate this behavior. Irino and Patterson in [5] have pointed out that our auditory system can do a kind of scale normalization (maybe using a Mellin transform) and in another paper [4] they present an application to this idea to their Auditory Image Model. Our objective is to study the hypothesis to apply the Mellin transform to the spectral envelope of the signal in order to achieve some kind of pitch, gender, age normalization of the pronounced vowels, and all this in an efficient (fast, or quasi-realtime) way. Related works have been presented in [13] and [11] to obtain normalized envelopes or spectra and in [9] and [10] using Mel-scale warping and scale-cepstral coefficients for the same purpose. In [8] an application of the scale-cepstrum for speech analysis has been presented and in [12] an implementation of STCC (scale-transform cepstrum coefficients) has been developed and compared to MFCC (Mel-scale cepstrum coefficients).

2. PITCH AND VOCAL TRACT IN VOWELS

From a temporal point of view, the signals of the same vowels pronounced by the same person, but with different pitch are not equal. But a more accurate analysis of the signal reveals that a single period of the vowel presents similarities with the single period of the other. The different part of the signals is the "gap" between periods. So

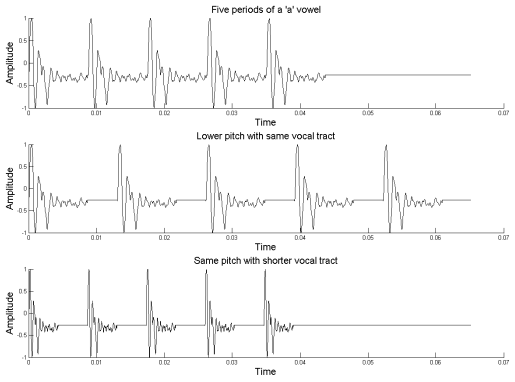

Figure 1. Five periods of an 'a' vowel (top), same vowel with a lower pitch, no change to vocal tract (middle), same vowel with same pitch and a shorter vocal tract (bottom).

we can build an artificial change of pitch taking a single period of a vowel, add zeros in tail and replicate this structure. For example you can get a new vowel pronunciation with a lower pitch with this modification (see figure 1).

Different ages and different genders can be studies using the ideas presented in [4]. Basically, the vocal tract of every person can be modelled with a lossless acoustic tube, so this tube is different only in length between different persons. The impulse responses of those tubes are one the scaled version of the other. So, again, if we take a single period of a vowel and confront it with another person (with a different length of the vocal tract) single period of the same vowel we should get two scaled versions of the same signal (see figure 1).

Now, the problem with these two ideas is that, for each signal, we must extrapolate the single period, find its start point and end point, and compare (directly in the pitch modification case, and with a scale normalization for the varied vocal tract case) the signals. Finding the single periods can be very difficult (pitch detection problem, cnfr. [7]) and results can be bad because a single period is short (too few samples) and can be different from others (we need some kind of mean).

But there is another way and we want to investigate about this other solution. Instead of working directly in time domain we can work on frequency domain. This give us two advantages: the first is that we "loose" time, so we don't care about time shifts between signals and we can

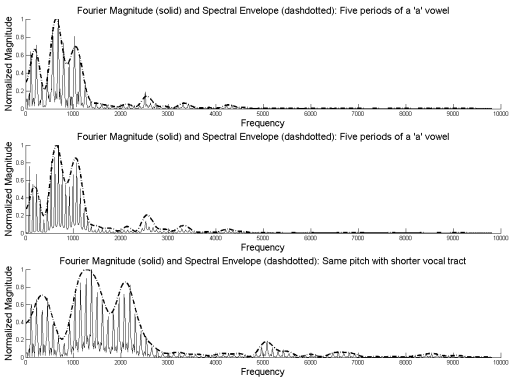

Figure 2. Fourier magnitude and spectral envelopes of the figure 1 signals. The envelopes estimations are very similar (top and middle) or in scale relation (same curve but more or less compressed, bottom).

avoid synchronization problems, the second is that if we work on the envelope we can avoid the signal preprocessing (find the single period, the start point, end point and compute some kind of mean).

So, the idea is to take the spectrum of the signal (only positive frequency [1], e.g. work on analytic signal), extract the envelope [2] and apply the scale transform to the envelope. Since the envelope of the discussed signals stays the same (with a different compression factor for different length vocal tracts, see figure 2), making a scale transform gives us a magnitude distribution identical for all this signals. In this way, in theory, we have a system to recognize the same vowel pronounced by different people with different pitches.

Naturally, this is a simplified model, because there are some differences that we discard (vocal tract geometry is slightly different between males and females, for example females have shorter pharynges in relation to their oral cavities [13]). Moreover, there is not a perfect constant compression factor between envelopes (authors of [10] suggest a Mel-scale warping instead a compression), so our results can be affected by these simplifications, but verifying that these reductions are not too heavy is part of this study.

3. THE SCALE AND MELLIN TRANSFORMS

The Mellin transform of a function f is defined as:

$$M_f(p) = \int_0^\infty f(t)\, t^{p-1}\, \mathrm{d}t\ , \tag{1}$$

where $p \in \mathbb{C}$ is the Mellin parameter. The scale transform [2] is a particular restriction of the Mellin transform on the vertical line $p = -jc + \frac{1}{2}$, with $c \in \mathbb{R}$. Thus, the scale transform is defined as:

$$D_f(c) = \frac{1}{\sqrt{2\pi}} \int_0^\infty f(t)\, \mathrm{e}^{(-jc-\frac{1}{2})\ln t}\, \mathrm{d}t. \tag{2}$$

The scale inverse transform is given by

$$f(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^\infty D_f(c)\, \mathrm{e}^{(jc-\frac{1}{2})\ln t}\, \mathrm{d}c. \tag{3}$$

The key property of the scale transform is the scale invariance. This means that if f is a function and g is a scaled version of f, the transform magnitude of both functions is the same. A scale modification is a compression or expansion of the time axis of the original function that preserves signal energy. Thus, a function $g(t)$ can be obtained with a scale modification from a function $f(t)$, if $g(t) = \sqrt{\alpha} f(\alpha t)$, with $\alpha \in \mathbb{R}^+$. When $\alpha < 1$ we get a scale expansion, when $\alpha > 1$ we get a scale compression. Given a scale modification with parameter α, the scale transforms of the original and scaled signals are related by

$$D_g(c) = \alpha^{jc} D_f(c). \tag{4}$$

This property derives from a similar property of the Mellin transform. In fact, if $h(t) = f(\alpha t)$, then

$$M_h(p) = \alpha^{-p} M_f(p). \tag{5}$$

In both (4) and (5), scaling is reflected by a multiplicative factor for the transforms, and for (4) such factor reduces to a pure phase shift. So, the scale transform magnitude of the original signal and the scaled signals is the same.

$$|D_g(c)| = |D_f(c)|. \tag{6}$$

4. ALGORITHM

Following the theory we have built an algorithm for studying the applicability of this method in practice and evaluating its performance. The goal is to obtain an automatic and simple (few or none controls, no tuning, no active control) recognition system that can map the same vowels to the same distribution and different vowels to different distributions. The system can be viewed as a sequence of steps: the first step is the computation of the Fourier transform magnitude for positive frequencies (or the Fourier transform magnitude of the analytic signal).

The second step is to build the spectral envelope. There are different algorithms and ideas on how to extract a spectral envelope (Channel Vocoder, LPC, Cepstrum). For our purpose we chose the cepstrum method [1]. The cepstrum method allows the estimation of a spectral envelope starting from the Fourier transform of the signal. First the signal (or a frame of the signal) can be windowed with a Hanning, Hamming or Gaussian window, then the log of the Fourier transform magnitude is computed (real cepstrum), after that the inverse Fourier transform is calculated, weighted with a particular low pass window [6] and

[1] The whole support is redundant and can cause a zero-in-head problem (reintroduction of a shift/synchronism problem that destroys the scale relation between envelopes).

[2] We cannot use directly the spectra because, in general they are not in scale relation, but their envelope are.

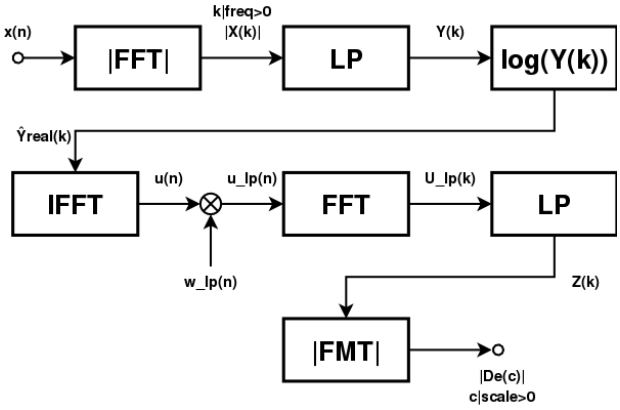

Figure 3. Algorithm graphical description.

finally another Fourier transform is applied. In our implementation we don't apply the windowing (for these experiments we work directly with already windowed signals chunks) and we compute the envelope on the positive frequencies only. The only parameter that we need to "tune" is the cut quefrency [3] value (the low pass cut value). For our purpose the envelope must be smooth enough to "absorb" the slight differences between same vowels, but not as smooth as to have different vowels mapped to similar envelopes. To achieve a good tradeoff, we applied a low pass filtering of the spectrum (before computing the real cepstrum) and another low pass filtering applied to the spectral envelope. Low pass filters cut frequencies are computed using the quefrency value, so we still have only one tuning parameter.

The third step is the scale transform of the spectral envelope. In theory, same vowels have same spectral envelope (or same envelope but with different compression or scale factor), so using the scale transform we should obtain the same magnitude distributions. Using the fast Mellin transform [3] algorithm we can compute the scale magnitude quickly ($\mathcal{O}(n \ln^2 n)$, see below).

The fourth and last step is a normalization of the scale magnitude distribution for comparisons between signals with different energy, although we can make this normalization before computing the scale transform.

Like already said, the algorithm has a unique control, the quefrency parameter for the envelope estimation and this makes using this technique very straightforward. The major flaw in this method is the envelope estimation. In fact the envelope is not a well defined curve, and different techniques or parameters give us different results. Therefore, for getting sufficiently good results we need to compare signals of the same class (e.g. same sampling factor, same quantization, similar recording conditions, etc.).

The asymptotic complexity of the entire procedure is $\mathcal{O}(n \ln^2 n)$, where n is the number of samples of the audio signal. In fact, the first two steps are $\mathcal{O}(n \ln n)$, because we need to compute Fourier transforms (FFT), the third step is $\mathcal{O}(n \ln^2 n)$, because this is the computational complexity of the scale transform [3], and the last step is linear.

[3] The term "quefrency" is commonly used when referring to the independent variable of the cepstrum domain.

5. TESTS

For testing the applicability of the whole idea, we have used artificial and real vowels.
The artificial vowels have been built from a unique real 'a' vowel (a_canon_A1_T1_real).
A single period was extracted and replicated to build an artificial 'a' (same pitch, same vocal tract, a_canon_A1_T1). With the same period we built another 'a' with a lower pitch, as described in section 2 (zero padding to have a longer period, a_canon_A1_T15_low). Then other two versions have been created: one (a_canon_A2_T1_high) is the simulation of a vocal tract reduction (again, as described in section 2, obtained by compression of the single period and zero padding the period to have the same length of the original one) and the last version is a pitch and vocal tract modification at the same time (a_canon_A2_T3_cmpr_high, not shown in the figures, a_canon_A2_T1_high has an almost equal distribution).

The real vowels are a_canon (a_canon_A1_T1_real is the first 0.132 seconds of it), e_canon, i_canon, o_canon and u_canon.

As can be seen in figure 4, all the artificial vowels can be mapped to very similar distributions. The scale axis is zoomed in the $0 - 5$ interval because there differences or similarities can be appreciated, and the magnitude is normalized. Of course they are not perfectly identical, but results are encouraging, especially when comparing a real 'a' vs. the artificial 'a' (figure 5).

Real vowels distributions (figures 6 and 7) appear different (even if there's not dramatic differences, for example 'a' and 'u' present some similarity) so they are distinguibile from each other.

The next step will be to make comparisons with real sounds between different genders and pitch, with an introduction of other components (e.g. an algorithm for automatic computation of quefrency parameter) in the system to go deeper in the study of applicability of this method for vowel recognition. Moreover, some kind of clustering must be done to verify that all the vowels can be mapped in different sets.

6. CONCLUSIONS

In this paper we have analyzed the applicability of the Mellin/scale transform for vowels recognition, focusing on spectral envelope scale distribution. We have reached a first indication that this idea can be pursued. The algorithm implemented is not usable yet, because it does not provide us a clear-cut answer (e.g. this vowel is equal to that), and it needs spectral envelope tuning (quefrency parameter). An expansion of the algorithm should be possible using automatic tuning. The tests have been limited to two classes of vowels, artificial ('a') and real (all vowels)

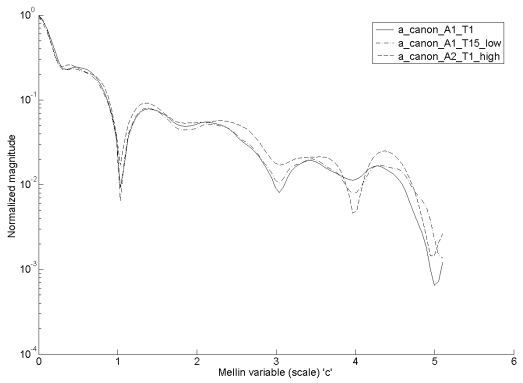

Figure 4. Log plot of the normalized magnitude of three envelope scale transforms. An ‘a’ vowel, a pitch lower ‘a’ and an ‘a’ pronounced by a shorter vocal tract. All the vowels are artificial.

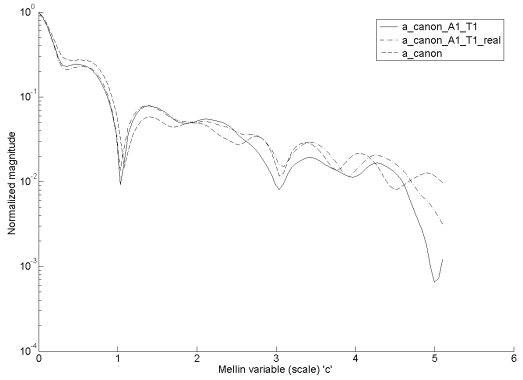

Figure 5. Log plot of the normalized magnitude of three envelope scale transforms. An artificial ‘a’ vowel, a real ‘a’ and a different length real ‘a’.

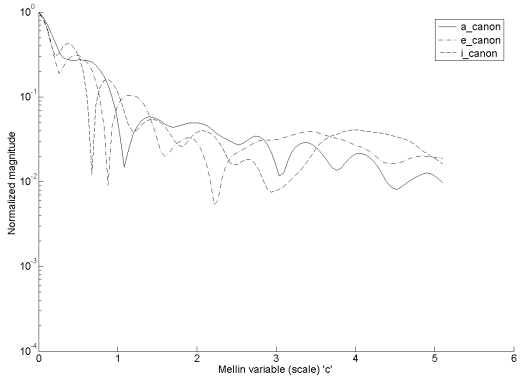

Figure 6. Log plot of the normalized magnitude of three envelope scale transforms. A real ‘a’ vowel, a real ‘e’ and a real ‘i’.

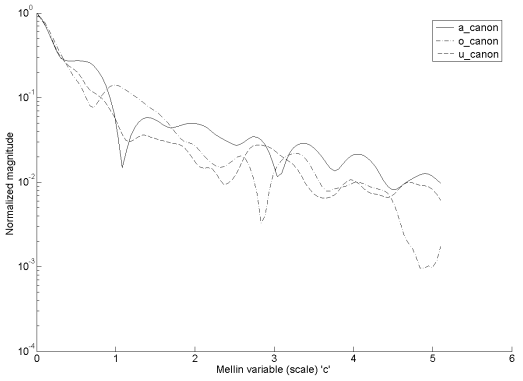

Figure 7. Log plot of the normalized magnitude of three envelope scale transforms. A real ‘a’ vowel, a real ‘o’ and a real ‘u’.

and can be further extended with real vowels only after the aforesaid algorithm modifications. The asymptotic complexity of the entire procedure is $\mathcal{O}(n \ln^2 n)$, so it shall be usable in realtime or quasi-realtime environment (depends on the signal length). The experiments have shown us that following the idea of a time-shift normalization (through Fourier transform and spectral envelope in particular) and a scale normalization (through scale transform) applied to audio signal can be pursued to make some kind of vowel recognition or normalization independent from who (age, gender) have pronounced the vowel and what pitch has been used.

7. REFERENCES

[1] D. Arfib, F. Keiler, and U. Zölzer. Source-filter processing. In U. Zölzer, editor, *Digital Audio Effects*, pages 299–372. John Wiley and Sons, Ltd., Chichester Sussex, UK, 2002.

[2] L. Cohen. The scale representation. *IEEE Trans. on signal processing*, 41(12):3275–3291, December 1993.

[3] A. De Sena and D. Rocchesso. A fast mellin transform with applications in dafx. In *Proc. of the 7th Int. Conference on Digital Audio Effects (DAFx'04)*, pages 65–69, October 2004. Naples, Italy, October 5-8.

[4] T. Irino and R. Patterson. Extracting size and shape information of sound source in an optimal auditory processing model. In *CASA workshop, IJCAI-99*, August 1999.

[5] T. Irino and R. Patterson. Segregating information about the size and the shape of the vocal tract using a time-domain auditory model: The stabilised wavelet-mellin transform. *Speech Communication*, 36(3):181–203, March 2002.

[6] A. V. Oppenheim and R. W. Schafer. *Digital Signal Processing*. Prentice-Hall, 1975.

[7] M. R. Schroeder. *Computer Speech*. Springer, 1999.

[8] S. Umesh, L. Cohen, N. Marinovic, and D. J. Nelson. Scale-transform in speech analysis. *IEEE Transactions on Speech and Audio Processing*, 7(1):40–45, January 1999.

[9] S. Umesh, L. Cohen, and D. Nelson. Frequency-warping and speaker-normalization. In *ICASSP-97, IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 983 – 986, April 1997.

[10] S. Umesh, L. Cohen, and D. Nelson. Frequency warping and the mel scale. *Signal Processing Letters*, 9(3):104 – 107, March 2002.

[11] S. Umesh, S. B. Kumar, M. K. Vinay, R. Sharma, and R. Sinha. A simple approach to non-uniform vowel normalization. In *ICASSP '02, IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 517 – 520, May 2002.

[12] S. Umesh, R. C. Rose, and S. Parthasarathy. Exploiting frequency-scaling invariance properties of the scale transform for automatic speech recognition. In *ICSLP-2000*, volume 1, pages 301 – 304, October 2000.

[13] H. Wakita. Normalization of vowels by vocal-tract length and its application to vowel identification. In *IEEE Transactions on Acoustics, Speech, and Signal Processing*, volume 25, pages 183 – 192, April 1977.