TOWARDS MORPHOLOGICAL SOUND DESCRIPTION USING SEGMENTAL MODELS

Julien Bloit,
IRCAM UMR CNRS-STMS,
Paris, France
julien.bloit@ircam.fr

Nicolas Rasamimanana,
IRCAM UMR CNRS-STMS,
Paris, France
nicolas.rasamimanana@ircam.fr

Frederic Bevilacqua,
IRCAM UMR CNRS-STMS,
Paris, France
frederic.bevilacqua@ircam.fr

ABSTRACT

We present an approach to model the temporal evolution of audio descriptors using Segmental Models (SMs). This method allows to segment a signal as a sequence of primitives, constituted by a set of trajectories defined by the user. This allows one to explicitly model the time duration of primitives, and to take into account the time dependence between successive signal frames, contrary to standard Hidden Markov Models. We applied this approach to a database of violin playing. Various types of glissando and dynamics variations were specifically recorded. Our results shows that our approach using Segmental Models provides a segmentation that can be easily interpreted. Quantitatively, the Segmental Models performed better than standard implementation of Hidden Markow Models.

1. INTRODUCTION

One way of producing innovative music is to add complex sounds to the composer's vocabulary. We can think of various examples such as noise machines of the italian Futurists[1], sounds produced from electronical devices as well as extended playing techniques on traditionnal instruments [1, 2, 3]. Along a single sound event, complexity can be introduced by modulating pitchness, the timbre enveloppe, granularity etc. In such cases, an elementary sound can not be described only with steady values for pitch, timbre, duration and intensity values, which is the modeling assumption behind most systems designed for western music transcription.

If one wants to describe such complex notes, it is desirable to seek for existing sound ontologies [4, 5]. We pursue ideas from previous works [6, 7] where authors aimed to implement ideas from Pierre Schaeffer's description of sounds[8]. One could roughly describe his system as the representation of complex notes as characteristic temporal profiles on perceptual dimensions. In these works, the authors designed temporal features in order to fit Schaeffer's morpholgical sound descriptions. There are several limitations in these previous attempts that we try to overcome in the present work. The first comes from considering global features on an isolated portion of sound as opposed to instantaneous features. This is suboptimal from a statistical learning standpoint. Defining sub-units would allow to limit the number of models [9]. Moreover, the combination of sub-units can lead to more expressive models, as the use of a limited set of phonemes allow the modeling of a high number of words.

In the proposed approach, we try to overcome these limitations by proposing a statistical framework to explicitly model audio descriptor trajectories. The modeling philosophy consists in taking maximum advantage of our prior knowledge that data can be viewed as trajectories, so that subsequents observations are strongly correlated. This segmental approach already used for handwriting modeling in [10] has proven to be a good solution when only little training data is available. Furthermore, explicitly modeling the duration has shown to increase robustness to noisy conditions [11]. The statistical framework is based on Segmental Models (SMs). SMs are a generalization of Hidden Markov Models (HMM) [12] that address three principal HMM limitations: 1) weak duration modelling, 2) assumption of conditional independence of observations given the state sequence and 3) the restrictions on feature extraction imposed by frame-based observations [13]. To the contrary, SMs provide explicit state duration distributions, explicit correlation models and use segmental rather than frame-based features.

The paper is structured as follows. In section 2 we introduce the formalism of SMs and how we adapt it to audio descriptors. In section 3, we present an experimental set up to validate our approach. We finally present the results of a classification task, and give perspective for future studies in sections 4 and 5.

2. SEGMENTAL MODELS

In this section, we present some key points of the SM formalism to model time dynamics. This modelling is based on a set of curve primitives that we introduce here. We also describe the decoding process that permits to segment a signal into a sequence of curve primitives.

2.1. Model Description

The SM formalism addresses two aspects that are particularly essential for our approach. We briefly review these two points and we invite the reader to refer to [13] for a more in-depth presentation of SMs. First, contrary to HMMs where observations are

[1] The Intonarumori (noise intoners) built by Russolo in the early 20th century.

assumed to be independent from each other, SMs directly model sequences of observations. Each state represents elementary curve shapes, also called primitives. This first property enables to fully consider possible time dependence between successive signal frames thanks to the use of explicit curve shapes. The second property addresses duration modeling of states. In SMs, the time spent in each state is defined in a flexible way, using duration distributions. This permits to reflect that each curve shape possesses a characteristic duration length with some variability. Combined together, these two properties enables to consider curve primitives with possible amplitude and/or time deformations, which grants a flexible framework for the modelling of shapes. SMs have shown successful in data mining to identify patterns in time series [14], or to provide a higher level representation in handwriting recognition tasks [10]. We here extend the idea to model time shapes in audio feature curves.

We represented on Figure 1 the general concept of the segmental approach applied to a monodimensional signal. We built an ergodic model where each state S_i is a predefined curve primitive: for each curve primitive, several duration lengths l_j are possible. This topology then enables to decompose the input signal into a sequence of primitives, each characterized by a duplet (S_i, l_j), using the decoding procedure presented in section 2.3.

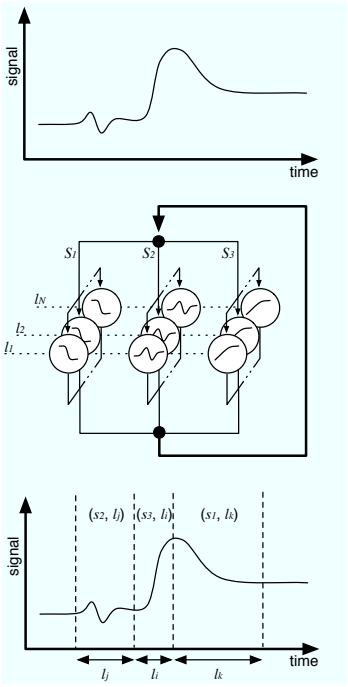

Figure 1: Model topology for the SM: each state represents a curve primitive S_i with possible duration lengths l_j. The model is fully connected. The decoding procedure then segments an input signal into a sequence of duplets ($primitives, lengths$).

2.2. Trajectory Models

From the model description, it appears that the choice made for the set of curve primitives is crucial. It not only conditions the obtained segmentation, but a judicious choice of primitives can additionally grant a level of interpretation on the signal decomposition. In this paper, we defined a set of primitives a priori as done in [15]: the primitives are segments with constant or weak curvature, with slopes equally distributed within $[-\pi/2; \pi/2]$.

A T-long trajectory is generated using an initial angle θ_{init}, a final angle θ_{final}, and the following linear interpolation:

$$\theta_t = \theta_{init} + \frac{n}{T-1}(\theta_{final} - \theta_{init}),\ with\ t = [0, T-1]$$

Varying T, we obtain different lengths of elementary trajectories. Varying θ_{init} and θ_{final}, we can set the main segment angle. A set of nine such elementary models is illustrated on Figure 2a. Each segment represents an archetype building block for a feature curve, in the sense that it is built upon the idea that any feature curve could be roughly described as a concatenation of successive segments with various durations.

The reason for choosing this set is partly inspired by the work in [16] where the author compared an analogous predefined set to a more specific one, learned from several handwriting datasets, and found that the predefined one were generic enough to account for any handwriting curve. We adapted it using only segments in the x-positive plane. Although quite basic, these curve primitives can capture possible trends of signal, typically stationary, going up or down. In addition, these features actually match aspects of the sound typology proposed by Schaeffer [8]. More advanced primitives can also be defined, in particular primitives with more specific curve shapes.

Another important aspect in the modeling deals with the choice of a set of possible duration lengths for the primitives. This set actually controls the time deformations that each primitive can assume and parallely defines a temporal granularity.

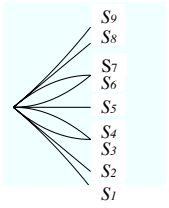

Figure 2: Set of nine curve primitives

2.3. Decoding

The decoding of the ergodic model yields to a segmentation of an input signal into the chosen primitives. We perform this step on the basis of a *maximum a posteriori* likelihood, with a 3D Viterbi procedure [13].

For an observed input signal $x_0...x_t$, we compute the corresponding sequence of angles $\hat{\theta}_1...\hat{\theta}_t$ to be invariant to possible curve offsets. We use the following formula :

$$\theta_t = \arctan((x_t - x_{t-1}) * fr) \quad (1)$$

where fr is the input signal's frame-rate.

Assuming a white gaussian noise b_t with variance σ for the observations, we get $\hat{\theta}_t = \theta_t + b_t$. The likelihood of the primitive S_k with respect to the observed sequence of angles $\hat{\theta}_1...\hat{\theta}_t$ is

approximated, as done in [10]:

$$-logp(\hat{\theta}_1 ... \hat{\theta}_t | t, S_k) = \frac{1}{2} \sum_{i=1}^{t} \frac{(\hat{\theta}_i - \theta_i)^2}{\sigma^2} \quad (2)$$

where θ_i is an element in the sequence of angles for S_k.

For an observed sequence of angles $\hat{\theta}_1 ... \hat{\theta}_T$, the decoding is based on $\delta_t(j)$, the log probability of the most likely sequence of elementary trajectories ending with trajectory label j, at time t:

$$\delta_t(j) = \max_{i=1,...,M} \max_{l \in \mathcal{L}} \delta_{t-l}(i)\, a_{ij}\, p_j(l)\, p(\hat{\theta}_{t-l+1} ... \hat{\theta}_t | l, S_j) \quad (3)$$

where a_{ij} is the transition probability from state S_i to S_j, M is the number of elementary trajectories, $\mathcal{L}$ is the set of possible duration lengths, and $p_j(l)$ the probability to stay in state S_j during l successive observations.

Choosing the maximum posterior probability path yields to two N-long sequences: S_1^N and l_1^N, where N is the number of states in the path. These two sequences actually give a representation of the input signal temporally decomposed on the set of primitives. Given our choice of curve primitives, this decomposition directly informs us of the signal trends over successive time ranges.

3. EXPERIMENTS

The approach was evaluated on a set of violin contemporary playing techniques. We describe here the datasets, the chosen audio description and the evaluation procedure.

3.1. Dataset

We specifically recorded data to carry out an evaluation of our approach. The music material involved various pitch and intensity profiles. To do so, we defined a musical vocabulary (see Figure 3a) composed of two pitch profiles (*upward glissando, downward glissando* and three intensity profiles (*crescendo, decrescendo, sforzando*, refered as $p_{1,2}$ and $i_{1,2,3}$ respectively. This vocabulary was chosen for the strong intrinsic temporal evolutions of its elements. *Crescendi* (resp. *glissandi*) consist in continuously progressing from one intensity level (resp. pitch) to another. $Sforzando$ consists in a step-like intensity profile with a louder part at the beginning. We generated short music sketches out of this vocabulary, by random combination of the vocabulary's elements with random pitches. Each sketch is a four-beat score, each beat being a combination of one intensity profile and one pitch profile. Moreover, no global dynamic levels were imposed, only *crescendi* and *decrescendi*. Figure 3b shows one example of a generated music sketch.

We automatically generated 43 sketches involving random proportions of pitch and intensity profiles. The generated scores were interpreted by a violin player at a given tempo of 60 bpm. Sound was recorded at 44100 Hz, and sliced into 46.4 msec windows, every 5.8 msec, yielding an approximate frame rate $fr = 172$ Hz.

3.2. Audio Features

We extracted two sound descriptors, highly correlated to the musical dimensions of pitch and intensity involved in our data, namely fundamental frequency [17] and loudness [18]. In order for the considered pitch profiles to be shift-invariant along the frequency axis, fundamental frequency was mapped from *Hertz* to a logarithmic scale (*cents*). The descriptor sequences were normalised within the $[0, 1]$ interval, using the possible violin ranges in pitch ($190Hz$ to $4400Hz$) and intensity ($0.01\ \ Sones$ to $15\ \ Sones$). Subsequently, these values are converted to angle sequences with Equation 1.

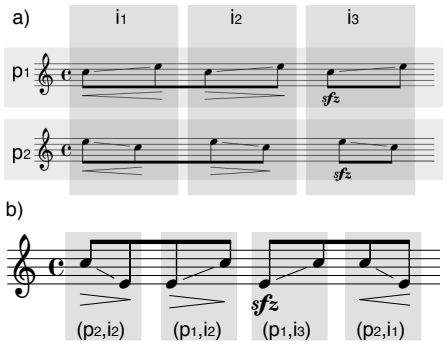

Figure 3: Pitch and intensity vocabulary elements (a) and sketch example generated from the combination of pitch and intensity profiles (b). Sketches were performed on a violin.

3.3. Evaluation Method

To assess our approach, we carried out a classification task on the vocabulary elements defined in section 3.1. The tasks can be identified as:

- task *T1*: classify the *upward* and *downward glissando* pitch profiles
- task *T2*: classify the *crescendo, decrescendo* and *sforzando* intensity profiles

The audio feature computation on each class element yields to a set of pitch and loudness values on which we separately ran a 3D Viterbi decoding. The output sequences of primitives and associated durations are then fed into a higher level HMM to constitute models of each vocabulary elements. This step is similar to the higher-level stage performed in [10] and can be seen as a way to agglomerate constitutive sub-units (i.e. the user-defined primitives) into larger semantic units. For this higher level HMM, we chose a 3-state left-right topology with a 2-dimensional Gaussian model and diagonal covariance to account for state indices and segment duration lengths.

The classification task was performed as follows. A typical train/test round consisted in training the higher-level model on a randomly picked 70% of the data, and testing on the remaining 30%. To evaluate a model on a given task, we ran each train/test round ten times in a row and averaged the classification scores on each test set. Training the models was done with conventional EM learning [19] using HTK [20] with simple left-right models. Classification scores were computed as the mean of diagonal terms on the normalised confusion matrix.

As a reference, we performed the same classification tasks with a HMM directly operating on the audio frames. We used the same 3-state left-right topology and train/test procedures, with a 1-dimensional Gaussian model to account for an incoming angle sequence $\hat{\theta}_1 ... \hat{\theta}_t$. The experiment protocol is summed up on Figure 4

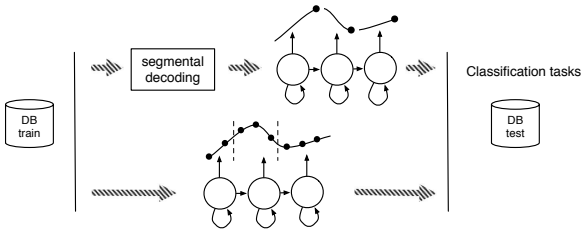

Figure 4: Experimental set overview. A left-right HMM with three states is trained with segmented observations output from the segmental decoding procedure. Classification results are compared to a left-right HMM operating on audio frames.

4. RESULTS

In this section, we first give a qualitative result intending to illustrate a typical output of the segmental decoding layer. We then give quantitative results on the classification tasks.

4.1. Segmentation results

The segmentation was performed using the primitives presented in section 2.2. We defined the set of possible duration lengths with values linearly taken between $230ms$ (roughly corresponding to short violin note) and $2.6s$ (several notes). The curves on Figure 5 show the resulting segmentation for one example of pitch profile and loudness profile.

We can see that on this example, the *glissando* is composed of three phases, i.e. flat pitch then increasing pitch and again flat pitch. This description in itself is quite informative on the violinist playing as we are able to see the details of his performance on this vocabulary element: in this example the pitch increasing phase was relatively short with two well defined flat phases. On the loudness profile, we can see that the *crescendo* is composed of a linearly increasing phase during most of the time before a rapid release.

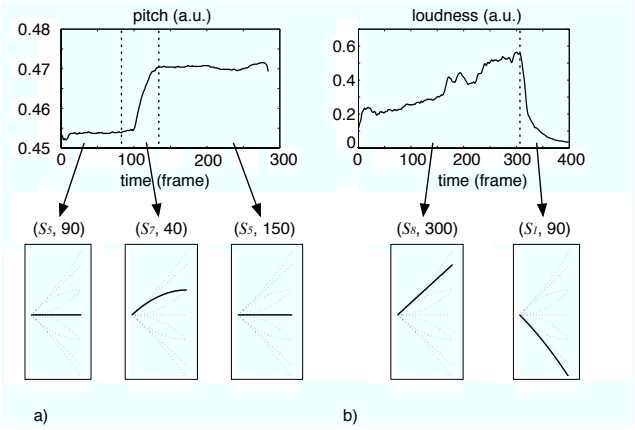

Figure 5: Segmentation results on two profile classes from DB1. a) shows a pitch profile for an *upward glissando* class. b) shows an intensity profile for a *crescendo* class. Below each feature curve, the sequence of primitive labels and durations is reported. In each box, the shape of the corresponding symbolic representation of each primitive is printed.

4.2. Classification results

Classification scores for pitch profiles displayed in Figure 6 (task *T1*) show that the segmental approach performs significantly better than the baseline frame-based approach (median value at 92% versus 72%). Moreover, the results also show more consistency as their variability is much smaller in the segmental approach (interquartile of 7 versus interquartile of 18). For the loudness profiles (task *T2*), results appear to be relatively similar between the two approaches (median value around 77%). However, the segmental approach shows once again a narrower variability in classification (interquartile of 3 versus interquartile of 12).

We can get insight of these results by inspecting the learned models and how the data fits. Figure 7 shows an example of the learned models for the two pitch profiles, for the frame-based approach (a and c) as well as for the segmental approach (b and d). As one could expect for this classification task, the second state seems to be the most discriminating one. In the frame-based approach, the second-state Gaussians only differ by a slight difference of mean, and tend to overlap. In the segmental approach, these second-state Gaussians are much more distinct. Interestingly, looking at the graphs, the segment duration observations do not seem to add much more discriminative power to the model. When inspecting loudness models, no such clear contrast was observed between the two approaches on the Gaussian distributions. In both cases, data looked less unimodal, which questions the chosen topology.

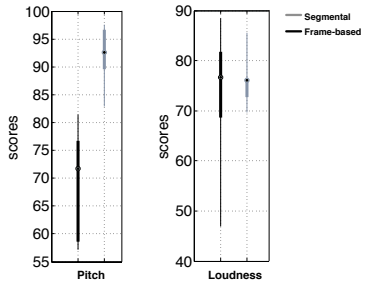

Figure 6: Classification scores on tasks *T1* and *T2*.

5. CONCLUSION AND PERSPECTIVES

We propose the use of Segmental Models to segment time curves of audio signals. The implementation we proposed was tested on two classification tasks using a database of violin contemporary playing. The segmental approach performed better than standard implementations of Hidden Markov Models in most cases. Importantly, Segmental Models overcome well-known limitations of HHMs, by explicitly modeling the time duration of primitives, and by taking into account the time dependence between successive signal frames. Future perspectives may adress the study of a realtime implementation on a data stream, using Viterbi extensions such as in [21]. The segmental approach performed well on a monophonic instrument in the context of contemporary music, however we believe that this approach can be easily extended to broader situations. In particular, we are investigating the use of more complex curve primitives to directly address specific sound components. Besides, we are also currently extending the ap-

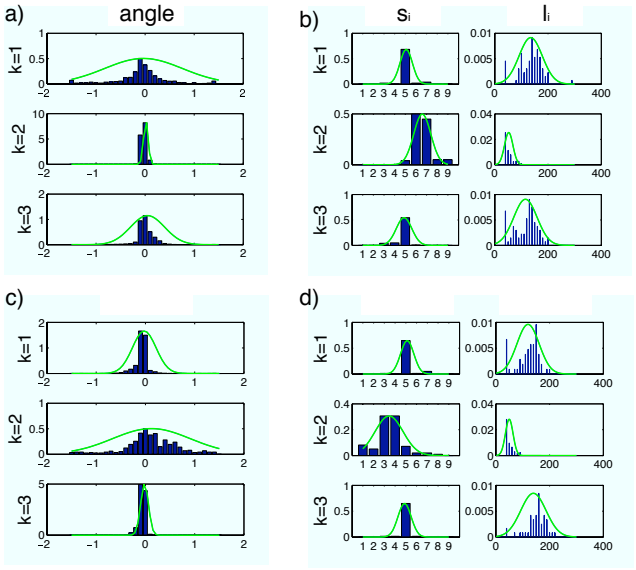

Figure 7: Observation densities and state aligned data for learned models for the two pitch classes. k stands for the state number of the model. a) states of the frame-based HMM for class p_1. b) states of the segmental HMM for class p_1 c) states of the frame-based HMM for class p_2 d) states of the segmental HMM for class p_2

proach to multidimensional features that could include other modalities like mouvement data.

6. ACKNOWLEDGMENTS

We would like to acknowledge Thierry Artières, Norbert Schnell and Xavier Rodet for fruitful discussions. This work has been partially supported by the ANR project Interlude.

7. REFERENCES

[1] M. Möller, "New sounds for flute," Available at http://www.sfz.se/flutetech, accessed April 05, 2009.

[2] F. Bevilacqua, N. Rasamimanana, E. Fléty, S. Lemouton, and F. Baschet, "The augmented violin project: research, composition and performance report," in *Proceedings of the International Conference on New Interfaces for Musical Expression (NIME)*, 2006.

[3] M. Kaltenecker, *Avec Helmut Lachenmann*, Van Dieren, Paris, 2001.

[4] M. Chion, *Guide des objets sonores*, INA/GRM, Buchet/Chastel, 1983.

[5] D. Smalley, "Spectromorphology : Explaining sound-shapes," *Organised Sound*, vol. 2, pp. 107–126, 1997.

[6] G. Peeters and E. Deruty, "Automatic morphological description of sounds," in *Acoustics 08*, Paris, 2008, SFA.

[7] J. Ricard and P. Herrera, "Morphological sound description computational model and usability evaluation," in *AES convention*, 2004.

[8] P. Schaeffer, *Traité des objets musicaux*, Seuil, 1966.

[9] L. Rabiner and B.H. Juang, *Fundamentals of speech recognition*, Prentice-Hall, Englewood Cliffs, 1993.

[10] T. Artières, S. Marukatat, and P. Gallinari, "Online handwritten shape recognition using segmental hidden markov models," *IEEE transactions on pattern analysis and machine intelligence*, vol. 29, pp. 205–217, 2007.

[11] A.C. Morris, S. Payne, and H. Bourlard, "Low cost duration modelling for noise robust speech recognition," in *ICSLP*, 2002, pp. 1025–1028.

[12] L. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," in *Proceedings of the IEEE*, 1989.

[13] M. Ostendorf, V. Digalakis, and O. A. Kimball, "From hmms to segment models: a unified view of stochastic modeling for speech recognition," *IEEE Trans. on Speech and Audio Processing*, vol. 4, pp. 360–378, 1996.

[14] X. Ge and P. Smyth, "Deformable markov model templates for time-series pattern matching," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2000, pp. 81 – 90.

[15] T. Artières and P. Gallinari, "Stroke level hmms for online handwriting recognition," in *Proceedings of the International Workshop Frontiers in Handwriting Recognition*, 2002, pp. 227 – 232.

[16] S. Marukatat, *Une approche générique pour la reconnaissance de signaux écrits en ligne. A generic approach to online handwriting recognition.*, Ph.D. thesis, Université Paris 6, 2004.

[17] A. de Cheveigné and H. Kawahara, "Yin, a fundamental frequency estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917 – 1930, 2002.

[18] B. C. J. Moore, B. R. Glasberg, and T. Baer, "A model for the prediction of thresholds, loudness, and partial loudness," *Journal of the Audio Engineering Society*, vol. 45, no. 4, pp. 224 – 240, 1997.

[19] Jeff A. Bilmes, "A gentle tutorial on the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models," Tech. Rep., U.C. Berkeley, 1997.

[20] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X.A. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book (for HTK Version 3.4)*, Cambridge University Engineering Department, 2006.

[21] J. Bloit and X. Rodet, "Short-time viterbi for online hmm decoding : evaluation on a real-time phone recognition task," in *ICASSP*, Las Vegas, 2008.

[22] G. Gravier, G. Potamianos, and C. Neti, "Asynchrony modeling for audio-visual speech recognition," in *Human Language Technology Conference (HLT)*, 2002.

[23] G. Peeters, "A large set of audio features for sound description (similarity and classification) in the cuidado project," Tech. Rep., IRCAM, 2004.

[24] J. O. Ramsay and B. W. Silverman, *Functiona Data Analysis*, New York: Springer-Verlag, 1997.

[25] P. Leveau, E. Vincent, G. Richard, and L. Daudet, "Instrument-specific harmonic atoms for mid-level music representation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 1, pp. 116 – 128, 2008.