An Auditory Model Based Transcriber of Vocal Queries

Tom De Mulder, Jean-Pierre Martens
ELIS, Ghent University
Sint-Pietersnieuwstraat 41
B-9000 Gent (Belgium)
tdmulder@elis.ugent.be

Micheline Lesaffre, Marc Leman
IPEM, Ghent University
Blandijnberg 2
B-9000 Gent (Belgium)
marc.leman@ugent.be

Bernard De Baets, Hans De Meyer
KERMIT, Ghent University
Coupure Links 653
B-9000 Gent (Belgium)
bdebaets@ugent.be

Abstract

In this paper a new auditory model-based transcriber of vocal melodic queries is presented. Our experiments show that the new system can transcribe queries with an accuracy between 76 % (whistling) and 85 % (singing with syllables), and that it outperforms four state-of-the-art systems it was compared with.

1 Introduction

Nowadays, all existing QBH systems seem to consist of two parts: (i) an acoustic front-end to transcribe the acoustic input into a note sequence, and (ii) a pattern matching back-end to search in a database for the musical piece best matching this sequence. This paper focuses on the development and evaluation of a new acoustic front-end.

2 A new acoustic front-end

The new front-end is an extension of the one previously described in (Clarisse 2002). The embedded auditory model now has two pitch extractors working in parallel: (i) AMPEX (see Van Immerseel 1992, Clarisse 2002) which performs a temporal analysis of the individual auditory nerve patterns, and (ii) SHS which performs an analysis of the auditory spectrum and which is inspired by the Sub-Harmonic Summation theory of Terhardt et al.(Terhardt 1982). Per 10 ms frame the auditory model generates a discrete auditory spectrum (40 channels) plus AMPEX and SHS (pitch,evidence) pairs.

2.1 A combination of two pitch extractors

If the pitch of a periodic signal is sufficiently low, the auditory nerve patterns in most auditory channels will exhibit periodic patterns emerging from interactions between harmonics of this pitch (Van Immerseel 1992). These patterns can be analyzed in the time domain. However, if the pitch gets higher, fewer channels exhibit periodic patterns originating from the pitch. On the other hand, if consecutive harmonics appear in different channels, they give rise to maxima in the auditory spectrum, and the pitch will emerge from the positions of these maxima. The latter calls for an analysis in the frequency domain.

At each frame n, our SHS pitch extractor generates a set $T_n = \{F_k(n), A_k(n)\}$ (with $k = 1, .., K_n$) of frequencies and amplitudes of tones that are presumed to be present in that frame. It does so in three steps: (a) search for salient maxima in the discrete auditory spectrum, (b) refine the positions of these maxima by parabolic fitting, (c) if the auditory spectrum in the vicinity of such a position resembles that caused by a pure tone, add the maximum position (converted to Hz) and amplitude to T_n. Once the tone sets T_{n-1}, T_n and T_{n+1} are available, they are used to compute a pitch estimate for frame n. This is accomplished as follows:

1. Select each $F_k(n)$ and its first five sub-harmonics as potential pitch candidates (provided they fall in the range from 350 to 4000 Hz).
2. For each candidate $F(n)$, compute its evidence as the weighted mean of the amplitudes of all the tones in $T_{n-1}..T_{n+1}$ that coincide with a harmonic of $F(n)$: frequencies F_1 and F_2 are said to coincide if $|F_1 - F_2|/|F_1 + F_2| < \epsilon_F$. If there is coincidence with harmonic i the weight is γ^i.
3. For the most evident pitch candidate F, recompute its frequency as a weighted mean: each tone in $T_{n-1}..T_{n+1}$ that coincides with a harmonic of F contributes with the appropriate sub-harmonic frequency, and with the tone amplitude as a weight.
4. Once the refined pitch F is computed, its final evidence is computed as the sum of the amplitudes of the tones in $T_{n-1}..T_{n+1}$ that coincide with a harmonic of F.

With respect to resolution it should be noted that since AMPEX is only searching for pitches below 400 Hz, it can achieve a resolution of 3 % with cochlear channel outputs sampled at 2.5 kHz. Similarly, SHS is only looking for pitches above 400 Hz and can therefore achieve the same resolution with a sampling of the auditory spectra at multiples of 0.5 bark.

2.2 An improved segmentation strategy

In our original front-end, the segmentation of a query into note segments and white spaces was mainly based on an analysis of the total energy pattern $E(n)$. Now we propose a multi-stage segmentation method that incorporates the two pitch extractors and that can cope better with legato, vibrato and tremolo. The different stages can be described as follows:

Pre-segmentation. In this stage, fully described in Clarisse 2002, candidate boundaries are generated at clear minima in $E(n)$ and anywhere $E(n)$ drops below a white space threshold.

Segment labeling. Every segment is labeled as a white space (WS), a low-frequency note (LF) or a high-frequency note (HF). If the maximum AMPEX evidence exceeds some threshold V_{min} (see Clarisse 2002) the label is LF, else, if the maximum SHS evidence exceeds some threshold V_{SHS}, the label is HF, else it is WS. Once the labeling is performed, a segmental pitch is computed for each note segment.

Boundary labeling. Candidate segment boundaries are marked as *reliable* or not. An unreliable boundary corresponds to a weak energy dip *between two notes with the same label*. The dip is characterized by a *dip fraction* f_{dip}. It is the ratio between the energy at the boundary and the minimum of the two surrounding energy maxima. A weak dip corresponds to a dip fraction $f_{dip} > \epsilon_{dip}$.

Boundary elimination. Unreliable boundaries are subjected to a more detailed analysis which takes into account f_{dip} and the difference ΔF_o (in semitones) between the segmental pitches of the two surrounding segments. Boundaries with a $\Delta F_o < a\, f_{dip} + d$ are eliminated. Experimental data revealed different best combinations (a_{LF}, d_{LF}) and (a_{HF}, d_{HF}) for the elimination of LF notes and HF notes respectively.

Legato processing. Some note boundaries are not marked by an energy dip, but by a pitch shift only. They are overlooked by the pre-segmentation, but they can be recovered by means of the following procedure applied to long (> 300 ms) notes:

- **Pitch stability analysis**. Determine for each frame the maximum interval to the right in which the minimum and maximum pitch still coincide (as defined before). The result of this analysis is a stable interval length pattern.
- **Stable interval detection**. From left to right, search for a maximum in the stable interval length pattern. If it exceeds 150 ms, mark the interval starting at that maximum as a stable pitch interval and move to the position right after that interval. Repeat this procedure on the remainder of the segment until the end of the segment is reached.
- **Legato decision**. In case of multiple stable intervals, consider the centers of the gaps between them as new boundaries and compute the pitches of the new segments.

3 Experimental results

The main goals of our experiments were: to assess the accuracy of the new front-end (MAMI) by comparing its transcriptions to manual transcriptions, and to compare this accuracy with that of other state-of-the-art systems like Solo Explorer (Rolland 1999), Ear Analyzer (Heinz 2003) and Akoff Composer.

3.1 Free parameters of MAMI

The free parameters of MAMI were set on the basis of experiments on a development data set. This gave the following results: $\gamma = 0.75$, $\epsilon_F = 0.025$, $V_{SHS} = 0.4$ times the maximal SHS evidence observed in the entire data set, $\epsilon_{dip} = 0.3$, $(a_{LF}, d_{LF}) = (3, -1.5)$ and $(a_{HF}, d_{HF}) = (2, 0)$. The most critical parameters are V_{SHS} and the two (a, d)-combinations.

3.2 Evaluation of different front-ends

The acoustic front-ends were tested on three types of queries: (a) singing with syllables, (b) singing with words and (c) whistling. The measures of discrepancy between generated and manual transcriptions are percent of note deletions+insertions, and total error, obtained by adding the percent of times a MIDI-rounded note difference of 2 or more semitones is observed. The results are listed in Table 1. The MAMI front-end clearly outperforms the other systems on all query types, but particularly on singing with words. For whistling, it is not that much better than Solo Explorer.

data set	error type	Evaluated acoustic FE			
		Akoff	Solo	Ear	MAMI
syllables	del+ins	82.1	20.1	15.9	10.4
(414 notes)	total error	97.8	23.0	20.5	15.5
words	del+ins	54.3	24.3	48.5	15.4
(657 notes)	total error	72.3	33.0	61.8	21.2
whistling	del+ins	73.1	24.7	35.0	20.8
(283 notes)	total error	79.8	28.9	37.5	23.6

Table 1: Evaluation of four front-ends on three test sets. The size of each test set (in notes) is mentioned between brackets.

4 Conclusions

The newly presented acoustic front-end can transcribe all types of vocal queries with an accuracy ranging from 76 % for whistling to 85% for singing with syllables. It clearly outperforms all other tested systems on all query types. It is also clear that most of the errors are segmentation errors, meaning that back-ends must be able to accommodate this type of errors.

5 Acknowledgments

This research was performed in the context of the Musical Audio Mining project which is funded by the Flemish Institute for the Promotion of the Scientific and Technical Research in Industry (grant 010035-GBOU). P.Y. Rolland, G. Raskinis and T. Heinz are acknowledged for granting permission to publish results obtained with Solo Explorer and Ear Analyzer.

6 References

1. Akoff Music Composer 2.0. Akoff Sound Lab. http://www.akoff.com.
2. Clarisse L., Martens J.P., Lesaffre M., De Baets B., De Meyer H., Leman M. (2002). "An auditory model based transcriber of singing sequences", Procs. ISMIR, 116-123.
3. Heinz T., Brückmann A. (2003) "Using a physiological ear model for automatic melody transcription and sound source recognition", AES 114th Convention (Amsterdam)
4. Rolland P.Y., Raskinis G., Ganascia J. (1999). "Musical content-based retrieval: an overview of the Melodiscov approach and system", Procs. ACM Multimedia, 81-84.
5. Terhardt E., Stoll G., Seewann M. (1982). "Algorithm for extraction of pitch and pitch salience for complex tonal signals", J. Acoust. Soc. Am. 71, 679-688.
6. Van Immerseel L., Martens J.P. (1992). "Pitch and voiced/unvoiced determination with an auditory model", J. Acoust. Soc. Am. 91, 3511-3526.